

# **Characterization of the Alpha Defensin Copy Number Variation in Humans**

Fayeza Fatima Khan

Thesis submitted to the University of Nottingham  
for the degree of Doctor of Philosophy

March 2012

# Acknowledgements

Foremost I would like to thank John for giving me the opportunity to work on this project. My success is down to his infallible mentoring. It has been a very educational and enjoyable time in JALA lab, a place I have looked forward to going to each day these last three years. Equally responsible for the great atmosphere and good times are my friends in the lab: Jess, Raquel, Suhaili, Somwang, Dannie, Tamsin, Sugandha, Holly, Dibo, Omniah, Laura and Ivan. I am especially thankful to Jess for helping me start up when I began, and continuous subsequent support, and to the alpha defensin team for sharing all things 'alpha'. For the Cystic Fibrosis analysis, I am grateful to Dr. Jane Davies for her collaboration.

I am indebted to the University of Karachi for their generous financial support for my PhD studies under HEC's scholarship program.

I must also thank all past teachers, but especially my Biology teacher at PAF College Lahore, Asim Butt. My base in biology (and love for it!) would certainly have been weaker without his direction.

Lastly, I will say thank you to all friends and family for their encouragement and love, especially my parents, Dilawer S. Khan and Arifa Khan, and my grandmother, Bilquees Jan.

# Contents

ABSTRACT .....	6
1. INTRODUCTION .....	7
1.1. Copy Number Variation in the Human Genome .....	7
1.2. Copy Number Assays .....	12
1.3. The Defensin Peptides .....	17
1.4. Human Alpha Defensin Genes and their CNV .....	23
1.5. Alpha Defensins and Disease: Effect of CNV .....	27
1.6. Aims of This Thesis .....	29
2. METHODS AND MATERIALS .....	31
2.1. Copy Number Measurement Assays .....	31
2.2. Copy Number Validating Assays (Allele Ratio Assays) .....	34
2.3. Capillary Electrophoresis .....	37
2.4. Assigning Copy Numbers .....	38
2.5. DEFA1A3 del Assay .....	39
2.6. Correlation with microarray data .....	39
2.7. Segregation Analysis .....	41
2.8. SNP Association with Copy Number Haplotypes .....	43
2.9. rs4300027 Genotyping Assay .....	43
2.10. DEFA4 Deletion Assay .....	45
2.11. PRT DEFTP1 and usDEL Assay .....	48
2.12. Upstream Sequence Replacement Assay .....	50
2.13. Protein Methods .....	52
2.14. Samples, General Reagents and Procedures .....	56

3.	<i>DEFA1A3</i> COPY NUMBERS .....	59
3.1.	Total Measured Copy Numbers .....	60
3.2.	Copy Number Measurement Quality .....	64
3.3.	DefHae3 and Indel-5 Assay .....	76
3.4.	Conclusion and Discussion .....	79
4.	<i>DEFA1A3</i> HAPLOTYPES .....	81
4.1.	Segregation Analysis of CEPH families.....	81
4.2.	Copy Number Haplotypes in Europeans.....	86
4.3.	Copy Number Haplotypes and SNPs in Europeans .....	92
4.4.	Copy Numbers and SNP Genotypes in Non-European HapMaps .....	98
4.5.	Upstream Replacement Polymorphism.....	102
4.6.	Conclusion and Discussion .....	108
5.	INTERROGATING GENOME WIDE ASSOCIATION STUDIES (GWAS).....	110
5.1.	SNPs as CNV Surrogates.....	111
5.2.	rs4300027 as a Surrogate for <i>DEFA1A3</i> Copy Number .....	112
5.3.	Investigating GWAS .....	113
5.4.	Conclusions and Discussion .....	117
6.	EXPRESSION OF <i>DEFA1A3</i> GENES .....	118
6.1.	Synthesis of DEFA1/2/3 peptides .....	118
6.2.	Neutrophil variability .....	119
6.3.	Determining Alpha Defensin Peptides in Neutrophil Extracts.....	121
6.4.	Acid Urea PAGE .....	125
6.5.	Quantifying Alpha Defensin Peptides (DEFA1/2/3).....	127
6.6.	Investigating a Role for Alpha Defensins in Cystic Fibrosis.....	128
6.7.	Conclusion and Discussion .....	137



7. DISCUSSION .....	139
7.1. Measurement of multiallelic CNVs .....	139
7.2. Evolution of <i>DEFA1A3</i> Haplotypes .....	141
7.3. <i>DEFA1A3</i> CNV and Clinical Phenotypes .....	143
8. Bibliography .....	146

# ABSTRACT

Copy number variation (CNV) has been acknowledged as an important contributor to variation in the human genome, and copy-variable regions harbouring genes are interesting subjects of research for their potential phenotypic effects. However, despite the extensive and continuing discovery of CNVs, multiallelic loci remain technically challenging to measure and study. In this thesis, a PCR-based measurement system for the tandemly repeated CNV that harbours the *DEFA1A3* locus has been developed. This locus can either have the *DEFA1* or the *DEFA3* gene in any given copy of the repeat; the two genes differ by a single nucleotide difference in the coding sequence. This CNV has previously been shown to vary from 4 to 11 copies in a sample of the UK population. The use of this measuring system has allowed a usefully accurate measure and some characterization of this CNV in Europeans, Asians (Chinese and Japanese) and African (Ibadan, Nigeria) samples from the HapMap project, agreeing with the previously found copy number range in Europeans and showing more variability in non-Europeans. Typing of three-generation CEPH families has allowed the inference of haplotype copy numbers from segregation analysis. Combining haplotype data for some CEPH HapMap samples that were part of these families with their SNP genotypes in the same LD block has shown copy number lineages that are surprisingly well-tagged by SNPs. SNP rs4300027 allows the division of copy number haplotypes into low (2 and 3-copy) and high (4 and 5-copy) groups in European HapMap samples, a result that has been corroborated in an independent set of European samples. The Asian and African HapMap samples have failed to show this particular association. However, Japanese samples show copy number lineages tagged by other SNPs, an association not observed in other populations. In the second part of the study, an attempt has been made to explore the phenotypic effects of this variable locus. Copy number-tagging SNPs have been used to investigate published GWAS for indirect signals of association with *DEFA1A3* copy number. Preliminary experiments for studying protein expression levels of *DEFA1* and *DEFA3* have been carried out and the possible role of this CNV as a modifier locus in Cystic Fibrosis disease severity has been investigated through direct typing of CF samples with clinical data, and indirectly through interrogating a CF GWAS.

# 1. INTRODUCTION

## 1.1. Copy Number Variation in the Human Genome

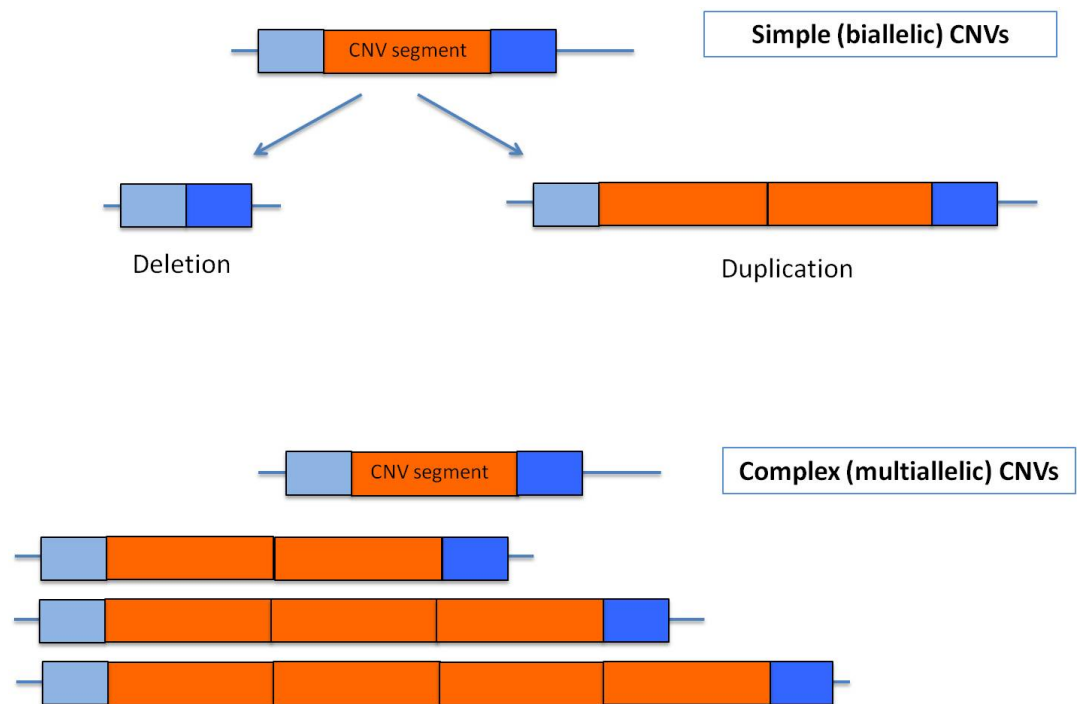
### Human Genomes Vary

The human genome is a dynamic entity. In a diploid cell, the nuclear genome is about 6 billion base pairs, packaged with the help of proteins into 23 pairs of chromosomes. It is subject to changes in its sequence and structure by cell processes, mainly replication of the genome prior to cell division and by extrinsic factors including chemical, physical and biological agents that are collectively labelled as mutagens. Most changes are thought to occur during replication. Although replication of the genome is a high-fidelity process, it is not always perfectly accurate and can result in errors. These errors can be of various kinds, but the mutation rate for the most common variation, single base substitution, is on average  $3 \times 10^{-8}$  mutations/nucleotide/generation (Xue *et al.*, 2009). Many changes in genes are not compatible with life, and thus most changes observed in human genomes are in the non-coding sequences and relatively few in the genes. However, because of obvious phenotypic changes, e.g. disease states, brought on by mutations in or around genes, these are disproportionately studied. Changes in genes may affect function by causing a decrease or complete loss, or may result in giving rise to new functional alleles or duplicated genes, or may not alter function. The sum of all the changes that have accumulated in the human genome, or were already present in the ancestors of humans, and are now observed in present day humans through various DNA technologies, is referred to as human genome variation. The most prevalent form of genomic variation is the change of single nucleotides, Single Nucleotide Polymorphisms (SNPs). About one SNP per 300 bp is observed. VNTRs (variable number of tandem repeats) like mini- and microsatellites, are less frequent with one per few kilobases, but are highly polymorphic with several kinds of alleles for any given locus. Small insertion/deletions (indels) are also about as frequent as microsatellites, but like the majority of SNPs, only biallelic. The most recently appreciated kind of genomic variation is structural variation, of which copy number variation is the most common.

## Copy Number Variation

Variation in copy number of human genomic segments containing genes has been observed for several decades. Genetic analyses of thalassaemia patients showed the deletion of alpha globin genes which are present in four copies in the general population instead of two copies per diploid genome. Similarly, the genetic analysis of Rh-negative versus Rh-positive individuals showed that the Rh-negative phenotype results from deletion of both copies of the *RHD* gene (Colin *et al.*, 1991). Such few examples of gene deletions/duplications which affected single-gene phenotypes and large, microscopically visible structural genomic changes were known before the era of human genome sequencing. With the availability of a reference human genome sequence (de Jong *et al.*, 2001) and subsequent development of genomic hybridization and SNP arrays, the last decade saw the discovery of the widespread copy number variation in the human genome (Iafrate *et al.*, 2004, Sebat *et al.*, 2004). Copy number variation is a kind of structural variation defined as the deletion or duplication of 1 kb or larger DNA segments such that they can be present in varying number of copies in contrast to the usual two copies for most of the genome in a diploid cell. Such variable DNA segments are usually referred to as copy number variants (CNVs), but can also be called copy number polymorphisms (CNP) in the literature (Redon *et al.*, 2006). While these two terms are often used interchangeably, sometimes they are used to differentiate between rare and common variants, as by McCarroll *et al.* (McCarroll *et al.*, 2007). Because of their size, these structural variations are thought to involve more DNA bases in variation between any two human genomes than SNPs (Redon *et al.*, 2006).

One way of classifying CNVs is on the basis of their origin and complexity: simple CNVs, which are usually biallelic and characterized by a single event of deletion or duplication such that all existing haplotypes are descendants of the original variant, and multiallelic or complex CNVs, which due to recurrent changes have several haplotypic structures, such that the same copy-number haplotype could have arisen more than once (Figure 1.1). This means that simple CNVs are generally found on the same haplotypic background on which they arose which allows their genotypes to be inferred indirectly by genotyping SNPs in LD with them (Hinds *et al.*, 2006). On the other hand, complex CNVs have undergone several rounds of copy number changes resulting in a complex spectrum of haplotypes in the population. Of course, a simple CNV could in due time become more complex with the addition of new haplotypes due to further structural changes, and all complex CNVs were at some point during their evolutionary history a simple duplication.



**Figure 1.1** A cartoon showing examples of simple and complex CNVs

So far CNVs have most commonly been discovered and studied using DNA array hybridization methods, where the loss or gain of DNA in one genome with respect to another is observable. Earlier arrays had DNA probes several kilobases long which allowed only a relatively rough estimation of CNV sizes. To better define CNV sizes and boundaries across the human genome, arrays with small DNA probes have since been designed and used. In one of the recent studies on genome-wide CNVs, 42 million element array with probe size approximately 70 base pairs has been employed (Conrad *et al.*, 2010). Key information that these arrays cannot give is the location of novel duplications. Now the development of next-generation sequencing technologies has made it possible to sequence thousands of human genomes in timescales of months rather than years, and such large-scale projects will not only help in defining the sequence basis of the already-discovered variants but also discover others.

## CNV Prevalence and Formation

According to the statistics from the Database of Genomic Variants (<http://projects.tcag.ca/variation>), there are about 16,000 CNVRs (CNV-containing regions) and about 67,000 CNVs deposited in the database. While some CNVs may still be missing, some could be documented twice because of

inaccurate characterization and others could be very rare, perhaps even one-off events. It is estimated that CNVs, ranging from 1 kb to several hundred kb, cover about 12% of the human genome (Stankiewicz *et al.*, 2010). As is true for all kinds of genomic variation, CNVs mostly include non-coding sequences. Few include exonic sequences. As Conrad *et al.*'s study has shown, most CNVs encompass only intergenic sequences, followed in frequency by intronic, promoter or exon sequences, and lastly whole genes. They are spread across all the genome and concentrated in those regions that already contain low-copy repeats, VNTRs, segmental duplications or transposable elements. The reason for this association is the predisposition of such sequences to result in structural changes that give rise to CNVs, and because of this predisposition, such regions (CNVRs) can harbor more than one CNV. Not only that, but to allow polymorphism to persist there should be no strong negative selection on copy number variation of such sequences (Nguyen *et al.*, 2008). Thus, common CNVs most likely represent genomic regions that can vary without negative consequences, rather than regions positively selected to be in higher copy numbers. Exceptions may exist. For example, it is theorized that higher copy numbers of the gene coding for salivary amylase have been positively selected for in populations depending upon starchy diets (Perry *et al.*, 2007).

Non-allelic homologous recombination between repeated sequences can result in the formation of CNVs. Other processes include non-homologous end joining (NHEJ) and microhomology-mediated break-induced repair (MMBIR) that occur during repair of double-strand breaks in the DNA. The evidence that these processes create CNVs comes from studying the breakpoints and sequences surrounding CNVs (Hastings *et al.*, 2009). These genomic changes can take place during meiosis or mitosis. The ones that occur at mitosis can either be restricted to one tissue type or present in several or all tissues depending upon the developmental stage and cell type in which they occur. For example, somatic mosaicism (Piotrowski *et al.*, 2008) and the observation of CNVs between monozygotic twins are the result of structural changes during mitosis (Bruder *et al.*, 2008).

How often are new CNVs formed? This question has been addressed for at least large CNVs (>100 kb) excluding regions of segmental duplications by a study looking at >700 transmissions (Itsara *et al.*, 2010). They found an unexpectedly low rate of  $1.2 \times 10^{-2}$  CNVs per transmission per generation compared to CNV prevalence known from higher resolution studies, indicating that purifying selection removes most large variants from the gene pool.

## CNVs and Phenotypes

Great interest lies in accurately mapping and genotyping CNVs because of their possible contribution to phenotypic variation, especially disease. There are two approaches to find a link between CNVs and phenotypes. One deals with known or already discovered CNVs, and involves studying one or several CNVs for their phenotypic effects. This is done either by directly correlating gene copy number with mRNA or protein expression levels (Stranger *et al.*, 2007) or indirectly by looking for copy number differences between individuals with phenotypic trait or disease state differences (case-control association studies) (Hollox *et al.*, 2008). This category would include commonly polymorphic CNVs that would not be strongly deleterious. The other approach is to search for CNVs across the genome (which are more likely to be rare or private variants) that are overrepresented in individuals with a particular trait or disease as compared to control individuals (Glessner *et al.*, 2009, Stefansson *et al.*, 2008).

Direct studies of the first kind are difficult to carry out in humans because it is difficult or impossible to access most living tissues from a cohort of living subjects for protein level studies. In addition, in cases of multiallelic CNVs, determining copy numbers accurately is another hurdle (discussed in Section 1.2). However, a study has been done using mice to look at effects of CNV on transcript levels (Henrichsen *et al.*, 2009). This has shown duplicated genes that have increased, decreased or unaffected transcript levels. It also showed that genes included in CNV segments were more likely to have more restricted, tissue-specific expression than genes never observed in CNVs. This was not an unexpected observation and is just another way of saying that the genes more essential to cell function are less likely to tolerate variation. An interesting example of a CNV not affecting mRNA dosage is of the green opsin genes (*OPN1MW*) in humans (Nathans *et al.*, 1986, Vollrath *et al.*, 1988). This gene can be repeated in tandem on the X chromosome and only the first copy in the array just downstream of the red opsin gene and promoter region is transcribed. If the first gene in the array has a loss-of-function mutation in a male, then no matter how many further copies follow, the person will be dichromatic (instead of trichromatic).

Several CNVs have been found associated with disease in humans. Large *de novo* deletions and duplications that are absent or very rare in the normal population are commonly found associated with neurodevelopmental defects like autism, schizophrenia and mental retardation (de Vries *et al.*, 2005, Sebat *et al.*, 2007, Walsh *et al.*, 2008). *De novo* CNVs also occur in phenotypically normal individuals, thus their presence alone is not strong evidence of

causality. Evidence that the *de novo* CNVs contribute to these disorders is either derived by considering the genes included in the CNV segment (e.g. neurodevelopmental genes) or previous observation of CNV of the same (or overlapping) region in similar phenotypes. In other cases they are experimentally verified in functional terms, as for example by gene knockouts in model organisms (Webber *et al.*, 2009). Low copy number (1 versus 2 or 3) of a large (~120 kb minimal) CNV was found associated with neuroblastoma risk (Diskin *et al.*, 2009). This CNV includes the gene NBPF23, member of a neuron-specific family of genes.

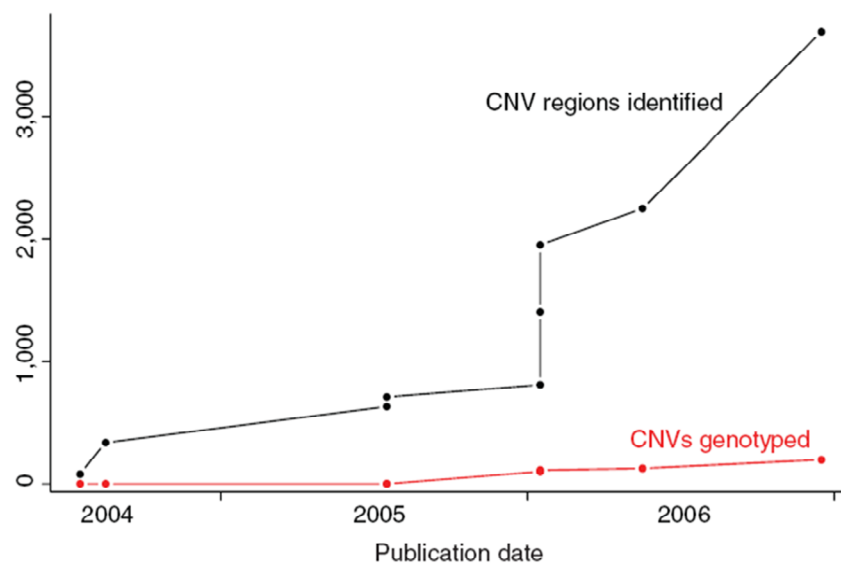
For CNVs that are common and highly polymorphic in the normal population, it would be natural to think that they do not confer strong negative effects (Itsara *et al.*, 2009). However, they could possibly contribute to multifactorial traits and are suspected to be part of the previously unaccounted for genomic variation (“missing heritability”) that contributes to a person’s predisposition to complex diseases like heart diseases, diabetes, cancer and Alzheimer’s, and variability in response to infectious or genetic diseases and drugs. For example, a common deletion upstream of the *IRGM* gene has been found to confer susceptibility to Crohn’s disease (McCarroll *et al.*, 2008). Altering gene dosage of *IRGM* was found to affect autophagy of internalized bacteria, a process considered to play a role in Crohn’s disease development. This deletion was found indirectly because of LD with a SNP that was found associated with Crohn’s disease but was not associated with any coding sequence variants (Parkes *et al.*, 2007). There is a dearth of such association studies for complex CNVs because of lack of accurate genotyping methods, and some that have been published, usually relying on real-time PCR, are bogged down by measurement inaccuracies (Shrestha *et al.*, 2009). The following sections discuss this issue in detail.

## **1.2. Copy Number Assays**

The discovery of CNVs has seen a rapid progress in the genomics era with the sequencing and availability of a more complete human genome and development of new methods like SNP arrays. Consequently, this particular kind of structural variation which had largely been unobserved previously became a strong contender for contribution to human phenotypic and disease variation. Alongside their discovery it became pertinent that methods to measure copy number of regions that are variable must be developed to further study their structure, mutation pattern and phenotypic consequences. However, discovery and genotyping methods did not develop at the same rate (Figure 1.2). Whereas simple CNVs involving a deletion or duplication and



having only two kinds of copy number haplotypes are easily measured, complex CNVs, like the alpha defensin CNV, that involve multiple copies and several different kinds of haplotypes require novel approaches. The challenge then is to accurately measure the number of copies in any given sample, without which further analysis is either impossible or the ability to get accurate results is greatly reduced. This does not include highly duplicated genes, like the ribosomal RNA genes, that range in the hundreds per genome. The use and power of accurately measuring *DEFA1A3* copy numbers for further analysis will be illustrated later on in this thesis.



**Figure 1.2** The difference in the number of CNV regions that have been identified in the literature and those that have been genotyped in available reference samples. Figure taken from McCarroll et al. (McCarroll *et al.*, 2007). The WTCCC CNV study from 2010 has genotyped >3000 CNVs (Craddock *et al.*, 2010)

Two studies have been published that undertook measuring the *DEFA1A3* CNV. One of these uses MAPH or Multiplex Amplifiable Probe Hybridization (Aldred *et al.*, 2005). In this study a set of seven samples was first determined for their *DEFA1A3* copy number by restriction digestion to cut out the repeat region containing all the repeats the sample contained, then separating and sizing the two haplotypes from each sample using pulsed-field gel electrophoresis and Southern blotting. These samples were then used to calibrate 110 other samples' MAPH measurement. **MAPH** (Armour *et al.*, 2000) uses probes for a CNV region(s) and reference sequences known to be 2-copy per genome that have a common set of sequences at their ends. Each probe is designed to vary in length from every other probe used. Genomic

DNAs of samples are denatured, immobilized on a membrane and the probes applied. The rationale is that the greater the number of copies of a sequence, the greater the amount of probe will be bound. The bound probes are then separated and amplified with a single set of primers which minimizes differences in amplification rates, and the products are analyzed by electrophoresis. This allows a measurement of ratios of test to reference sequence products that correlates with the copy number of test sequence. MAPH has been shown to be a robust method but a laborious and low-throughput one. The other study (Linzmeier *et al.*, 2005) used **Real-Time PCR**, which is an easy to perform and high-throughput method but the accuracy of results is unproven at higher copy number. The use of different primer pairs for test and reference loci may introduce a bias in amplification rates even with a very careful selection of primers. Even a small bias becomes amplified after several PCR rounds and renders this method unsuitable for accurate measurement. Studies have been published comparing CNV genotyping results from RT-PCR with other more robust methods, proving the generally low quality and inaccuracy of RT-PCR results (Aldhous *et al.*, 2010, Field *et al.*, 2009, Fode *et al.*, 2011). The two studies on alpha defensin CNV not only use different methods but also tested different populations; Aldred *et al.* reported a range of 4 to 11, with 7 copies being modal in 111 UK individuals, whereas Linzmeier *et al.* reported a range of 5 to 14, with 10 copies being modal, in 24 individuals from five different ethnic backgrounds (6 European American, 6 African American, 6 Asian American, 3 Mexican American and 3 Native American).

**SNP arrays** and **CGH-arrays** can also be used for copy number measurement, on the principle that the signal from a sequence correlates to its copy number in the sample. However, standard SNP arrays might not be very useful for common CNVs, as SNPs within common, complex CNV regions are not amenable to genotyping and have thus generally been excluded from SNP databases upon which the array designs are based. For CNV detection and genotyping special arrays have been designed and software created to detect and measure CNVs. With arrays using small interrogating probes (<100 bp) the resolution of structural variation has increased, or in other words, the breakpoints of CNV regions have become better defined. One study using aCGH with probes of the size range 70 bp for copy number determination failed to assign *DEFA1A3* copy number classes, among other CNVs, to individuals due presumably to failure of quality control (Craddock *et al.*, 2010). This may be due to the comparative nature of the study and failure to determine the copy number for the reference sample, which can be overcome if the reference sample's copy number has been previously determined by

other methods. However, arrays are appropriate for genome-wide CNV studies and not locus-specific studies.

With the development of second-generation **sequencing** technologies (Metzker, 2010), it is possible to sequence whole genomes at a great depth with short read lengths and use it for CNV measurement. The rationale is that the number of copies of a sequence in the genome will be represented proportionally in the number of sequenced reads. Thus the Illumina sequencing platform has been used by one study to detect, measure and characterize CNVs across the human genome in 159 individuals (Sudmant *et al.*, 2010). However, not all genomic regions are amenable to such an analysis due to bias in sequence capture and thus representation or even more importantly, due to complexity of structure e.g. short read lengths prevent correct mapping of repetitive sequences and thus although CNVs might be well-represented in the sequenced reads, they would remain intractable to further analysis. This method and using CGH-arrays are genome-wide approaches, but are still not as accurate and easy to perform as PRT, which is described below.

Other methods that are direct and locus-specific rather than comparative, like the **PFGE** experiment for *DEFA1A3* by Aldred *et al.* and **FISH**, are low-throughput. They may also be limited by the size of the CNV to be studied.

A locus-specific method that has been shown to combine both accuracy and high-throughput ability is the **Paralogue Ratio Test (PRT)** (Armour *et al.*, 2007). It is a simple PCR-based assay that uses a single pair of primers designed such that they amplify sequences at both the CNV locus and a reference locus which is two-copy per genome, with both sequences of different sizes. One of the primers used is labelled with a fluorescent dye. The products are separated and analyzed using a capillary electrophoresis system that detects the fluorescence from the primer and records the amounts of each product, which are displayed as peaks. Using the relevant software, the areas of the peaks corresponding to the test and reference loci are obtained and the ratio of test locus peak area to reference locus peak area is calculated. This ratio corresponds to the copy number of the test locus in a given sample, and the ratios of all samples give a linear relationship to their copy numbers. PRT uses a single pair of primers in a single reaction which reduces the effects of amplification bias or variable amplification efficiency between the loci. This is also a comparative method and requires the determination of copy number for some samples beforehand to be used as calibrators. It is important that whatever amplification bias between the test

and reference loci exists, it is the same constant value between calibrators and test samples. Quality of DNA can affect the rate of amplification. For example, one site may be a stronger protein-binding sequence than the other and the presence of residual proteins in the DNA samples could affect the access of primers in a biased manner between the test and reference loci. In a similar fashion, amplification bias can come from differences in test and reference loci sequences that affect the regional topology or melting temperature of the molecule and hence the access of primers to their complementary sites. Such differences in flanking sequences can also have the same effect. Although PRT uses similar sequences from test and reference loci as templates, this kind of differential can be present and its extent and effect is an important consideration in the development of a PRT. Despite these considerations for a PCR-based assay, PRT has been shown to be a high-throughput, accurate and robust copy number typing method and has been used for multiallelic CNVs (Walker *et al.*, 2009). Another limitation of this method is the requirement of the presence of paralogous sequences in the CNV region that allow such primers to be designed that amplify the CNV locus and a non-CNV reference locus, and completely exclude any other similar sequences in the genome.

**Table 1.1** Summary of the features of various CNV measuring methods when applied to complex CNVs.

Method	High-throughput	Accuracy	Applicable to all CNVs	Absolute Copy Number
RT-PCR	Yes	Usually low	Yes	No
MAPH	No	High	Yes	No
aCGH	No	Variable	No	No
2 <sup>nd</sup> Generation Sequencing	No	Variable	No	No
PRT	Yes	High	No	No
FISH	No	High	No	Yes
PFGE	No	High	No	Yes

Besides methods that measure copy number, there are methods that assist in copy number assignment by providing a ratio between the different copies of a CNV locus. These ratios are informative of the copy number on the rationale that a particular ratio is not possible from all copy numbers and thus excludes some copy numbers from consideration. For example, obtaining a ratio of 3 would mean a total copy number of 4 (3:1) or 8 (6:2) or higher multiples of 4. **Restriction Enzyme Digestion Variant Ratios** (REDVR) is one such method that uses multisite variants (MSVs: single nucleotide differences between the repeats of a CNV sequence) that are part of restriction enzyme recognition sites and digestion with the enzyme post-PCR amplification results in two kinds of products, the ratio of which is useful. Three such assays for the *DEFA1A3* CNV locus were used by Aldred *et al* (Aldred *et al.*, 2005). Similarly, small **indel polymorphisms** between repeats can be used. In this case just a simple PCR across the indel gives two kinds of products on the basis of size, and their ratio is useful. Similarly, the presence of a **microsatellite** in a CNV sequence allows the ratios of microsatellite alleles to be used as informative for total copy number of the CNV. Microsatellites are highly variable and can be very informative, but are prone to PCR slippage. The frequency of the minor allele in the case of indels and MSVs must be considered before the design of the assay to make sure it will be informative in most of the samples.

### 1.3. The Defensin Peptides

#### A class of AMPs

Defensins, as the name suggests, are antimicrobial peptides (AMPs) that provide immunity or defence against microbes. They are only one of several classes of AMPs found in plants, animals, fungi and bacteria, but the most prevalent one. AMPs are different from classical antibiotics in that they are products of protein synthesis from mRNA templates and so their structure is encoded in their respective genes. These are grouped into different classes on the basis of their structure. A characteristic structure of the defensins is the presence of two or more disulphide bridges, and thus four or more Cysteine residues, and a beta-sheet folding of the peptide. They are small (3 to 6 kDa), cationic peptides with much diversity in peptide sequence, antimicrobial properties and in the case of multicellular organisms, site of expression (Diamond *et al.*, 2009). Their pervasive prevalence and diversity reflect their importance in providing immunity to host organisms.

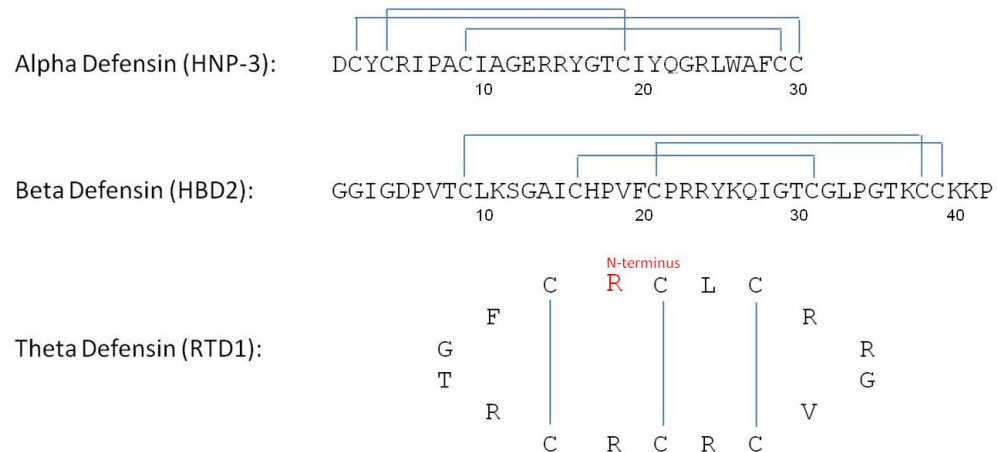
## Mode of Microbicidal Activity

Defensins, like other AMPs, exert their microbicidal effects mainly by inserting into and disrupting cell membranes, thereby causing the lysis of cells. They do this by forming pores that cause leakage of ions and larger molecules across the membrane, disrupting electrolytic balance and cell homeostasis (Wimley *et al.*, 1994). Their specificity for bacterial cells is due to the net anionic charge of the bacterial surfaces, whereas animal cells have relatively neutral surfaces. This mechanism makes them effective against enveloped viruses as well. Defensins are amphipathic, which means they possess both hydrophilic regions to interact with the charged surfaces of microbes and hydrophobic regions that allow their insertion in lipid membranes. Defensins have an overall cationic charge and the artificial replacement of the cationic residues with uncharged amino acids decreases their ability to kill bacteria (Zou *et al.*, 2007). The cationic charge allows their interaction with the anionic surface charge and non-polar residues facilitate insertion into the lipid bilayer. Whereas this appears to be the major mode of action, there is evidence that defensins, like alpha-defensin 1 from humans, can inhibit DNA, RNA and protein synthesis within bacterial cells (Brogden, 2005). The specific mode of action is thought to vary depending upon both the peptide and the microbe in question. Defensins have been found to be salt-sensitive with reduced activity at higher salt concentrations (Goldman *et al.*, 1997). Recently, it has been shown that a human beta-defensin peptide (hBD-1) that is constitutively expressed by the epithelial tissue has increased antimicrobial activity when its disulphide bonds are lost by reduction and that reduced hBD-1 is present in intestinal crypts and skin epithelia (Schroeder *et al.*, 2011). They also tested hBD-3 under reducing conditions for bactericidal activity and found reduced bacterial killing. Thus, whether there are other defensin peptides that may have enhanced microbicidal activity under reducing conditions or whether this mechanism is reserved for hBD-1 remains to be investigated.

## Defensin sub-classes

Several defensins have been identified in plants, invertebrates, like arthropods and molluscs, and in vertebrate animals, and are continuing to be discovered. In the literature these are classified according to the organism in which they are found like insect defensins, plant defensins etc. The defensins in vertebrates, especially mammals, on the other hand are much better studied, and these are further divided into three classes based on their structure (and hence phylogeny), rather than being named after their host organism.

Thus in vertebrates, on the basis of arrangement of disulphide bonds and the length of the peptide between them, defensins are sub-divided into alpha, beta and theta defensins. Alpha defensins have disulphide linkage between 1 and 6, 2 and 4, and 3 and 5 Cysteines, beta defensins have between 1 and 5, 2 and 4, and 3 and 6 Cysteines, and theta defensins are cyclic with no free end but contain the three disulphide bonds (Figure 1.3) (Ganz, 2003).



**Figure 1.3** Disulphide linkage characteristic of the three classes of primate defensins in their mature forms. The six Cysteines form disulphides in the order 1-6, 2-4, 3-5 for alpha defensins (30 amino acid mature Human Neutrophil Peptide-3 as example), 1-5, 2-6, 3-4 for beta defensins (42 amino acid mature Human Beta Defensin 2 as example), and a cyclic structure for theta defensins (18 amino acid Rhesus Theta Defensin as example). Figure modified from Ganz et al. (Ganz, 2003)

Theta defensins are formed by the ligation of two alpha-defensin related peptides (Tang et al 1999).

## Vertebrate Defensins

$\theta$ -defensins have only been identified in non-hominid primates, whereas  $\alpha$ -defensins have only been identified in Euarchontoglires (primates, rodents and lagomorphs). Defensins identified in all other vertebrates are  $\beta$ -defensins. Given the structural similarity of non-vertebrate and plant defensins to  $\beta$ -defensins, it seems obvious that these are the ancestral defensins that have given rise to  $\alpha$ -defensins in mammals. Theta-defensins have only been found to be active in some non-human primates, with pseudogenes present in other primates, including humans. They are thought to have arisen from mutations in the genes coding for alpha-defensins in Old World monkeys (Nguyen *et al.*, 2003). Thus these fast evolving peptides have also evolved differences in

tissue expression. The following table shows differences of expression sites between the defensins and between different mammals.

Species	Neutrophil defensins	Paneth cell defensins	Epithelial cell defensins
Human	$\alpha$	$\alpha$	$\alpha$ and $\beta$
Rhesus monkey	$\alpha$ and $\theta$	N.D.	$\beta$
Mouse	none	$\alpha$	$\alpha$ and $\beta$
Rat	$\alpha$	$\alpha$	$\beta$
Pig	Not detected in granule extracts	N.D.	$\beta$
Cow	$\beta$	none	$\beta$
Chicken	$\beta$	N.D.	$\beta$

N.D., not determined.

**Figure 1.4** Table taken from Ganz et al. (Ganz, 2003)

Studies on the amino acid sequences of these peptides and their secondary structures seem to show that the beta-defensins are evolutionarily more closely related to insect defensins than alpha-defensins, although it is not a perfectly clear picture (Hughes, 1999). This study by Hughes was prior to the availability of genome sequences of animals and he argues that although alpha and beta defensins have similar triple-stranded beta sheet structures, and their gene clusters are physically linked, both being present within a 100 kb of each other on chromosome 8, these are by no means conclusive grounds for inferring that they are related. Hughes has argued that the simultaneous and independent evolution of similar protein motifs occurs in evolutionarily non-related but functionally related genes, and close linkage, although common in closely related genes, is not an evidence of being related.

After the availability of genome sequences from various animal species, studies on the genomic clustering of defensin genes as well as comparison of amino acid and nucleotide sequences have led to the hypothesis that alpha-defensin genes originated from a beta-defensin gene precursor and that the theta-defensin gene originated from mutation in an alpha-defensin gene (Xiao *et al.*, 2004).

## Defensins in Humans

### $\beta$ -defensins

Seven  $\beta$ -defensins have been identified in humans at the protein level of expression, and a further 16 at the mRNA level. They are mainly expressed in



the epithelial cells of various tissues including respiratory tract, urinary tract, kidney, intestine, skin, conjunctiva, pancreas and testis. At least two  $\beta$ -defensins have been identified only in the epithelia of testis and epididymis, whereas the others have a range of tissue expression (Pazgier *et al.*, 2006). Their expression can be constitutional or induced. The control of their expression is not very well understood, but microorganisms or their structural components are known inducers and host molecules recognizing them are mediators of this induction (Kolls *et al.*, 2008).

### ***Function and Role in Psoriasis***

The function of  $\beta$ -defensins in infection and providing immunity is well documented. They have been shown to be effective *in vitro* against bacteria such as *E. coli* and *S. aureus* and fungi such as *C. albicans* (Vylkova *et al.*, 2007). Besides being antimicrobial, defensins also have other immunity-related functions. Beta-defensins have been shown to induce the production of cytokines, and also to act as cytokines themselves (Niyonsaba *et al.*, 2007). They induce proliferation and migration of keratinocytes, production of pro-inflammatory cytokines and thus are thought to be players in the wound healing process. In fact human  $\beta$ -defensin 2 (HBD-2) was first isolated from psoriatic lesions (Harder *et al.*, 1997) and whereas on the one hand they prevent the infection of these autoimmune lesions, a higher copy number of their genes has been associated with psoriasis (Hollox *et al.*, 2008), making them a contributor to the development of this disease possibly because of their ability to enhance inflammation.

### **$\alpha$ -defensins**

Six alpha defensins have been identified in humans. Their sizes range from 29 to 34 amino acids. Unlike the beta-defensins there is no discrepancy between the number of genes, mRNA transcripts and proteins identified for alpha-defensins except for alpha-defensin 2 (DEFA2). It is thought to be a proteolytic cleavage product of either or both of DEFA1 and DEFA3, as no gene/transcript has been found for it. DEFA1, 2, 3 and 4 (also known as Human Neutrophil Peptides or HNPs 1-4) are produced in the neutrophils, and DEFA5 and 6 (also known as Human Defensins or HDs 5 and 6) are produced in the Paneth cells of the small intestine. The production of DEFA1-4 takes place in promyelocytes, the precursor cells of neutrophils in the bone marrow, where they are packaged into azurophilic granules (Date *et al.*, 1994). No further synthesis of defensins takes place in mature neutrophils. These granules in the mature neutrophils fuse with phagocytised vacuoles containing pathogens, allowing defensins access to them.

(DEFA1)	AC <u>Y</u> <u>C</u> RI <u>P</u> AC <u>I</u> AG <u>E</u> RRYGT <u>C</u> IYQGRLWAF <u>C</u> <u>C</u>
(DEFA3)	DC <u>Y</u> <u>C</u> RI <u>P</u> AC <u>I</u> AG <u>E</u> RRYGT <u>C</u> IYQGRLWAF <u>C</u> <u>C</u>
(DEFA2)	<u>C</u> Y <u>C</u> RI <u>P</u> AC <u>I</u> AG <u>E</u> RRYGT <u>C</u> IYQGRLWAF <u>C</u> <u>C</u>
(DEFA4)	V <u>C</u> <u>S</u> <u>C</u> RLV <u>F</u> <u>C</u> RRTELRVGN <u>C</u> LIGGVSF <u>T</u> Y <u>C</u> <u>C</u> TRVD
(DEFA6)	FT <u>C</u> H <u>C</u> RR <u>S</u> <u>C</u> YSTEYSYGT <u>C</u> TVMGINHRF <u>C</u> <u>C</u> L
(DEFA5)	AT <u>C</u> <u>Y</u> <u>C</u> RTGR <u>C</u> ATRESLSGV <u>C</u> EISGRLYRL <u>C</u> <u>C</u> R

**Figure 1.5** Amino acid sequences of 6 human alpha defensins in their mature form with the six Cysteines in each peptide underlined

These defensins, as all others, are produced as inactive pre-prodefensins, which in the case of DEFA1 is 96 amino acids long. This includes an N-terminal signal peptide, an anionic propiece and the mature defensin. The signal peptide directs it to the endoplasmic reticulum from where it is packaged off into granules. This is cleaved to give the propeptide, 56 amino acids long. It is not known which enzyme(s) processes neutrophil defensins into their mature forms. The enteric alpha defensins were shown to be cleaved into their mature forms by Paneth cell trypsin (Ghosh *et al.*, 2002). The propeptide is then finally cleaved to give the mature defensin peptide (DEFA1), 30 amino acids long. Studies have found other intermediate-sized peptides that are larger than the mature peptide but of a different size than the pre-prodefensin and prodefensin. However, these precursor and intermediate peptides are found in minute quantities in the cell (0.25% of total defensin content) (Harwig *et al.*, 1992).

### ***Antimicrobial Properties***

Human alpha defensins are effective against bacteria, fungi and viruses, not just by direct microbicidal activity but also through the recruitment and modulation of other immune cells and functions. Each exhibits a different level of specificity for microbes. For example, DEFA4 and 5 are more potent against gram-negative bacteria like *E. coli*, whereas DEFA4 is least potent against gram-positive *S. aureus* whereas DEFA2 and DEFA5 are most effective (Ericksen *et al.*, 2005). DEFA1-3 have also been shown to neutralize toxins from *B. anthracis* and *C. difficile* (Giesemann *et al.*, 2008, Kim *et al.*, 2005). DEFA1-3 can inactivate Herpes Simplex virus and DEFA1 can inactivate influenza, vesicular stomatitis and cytomegalovirus (Daher *et al.*, 1986). This

direct inactivation is only seen against enveloped viruses, but indirect activity against non-enveloped viruses also exists via modulation of virus-infected cells (Klotman *et al.*, 2006). DEFA1 can also kill *T. cruzi*, the protozoan parasite causative of Chagas' disease (Madison *et al.*, 2007). Experiments whereby mice expressing HD-5 gene were compared with non-transgenic mice in their ability to protect from infection by Salmonella, showed the effectiveness of this defensin (Salzman *et al.*, 2003).

### ***Immunomodulatory Properties***

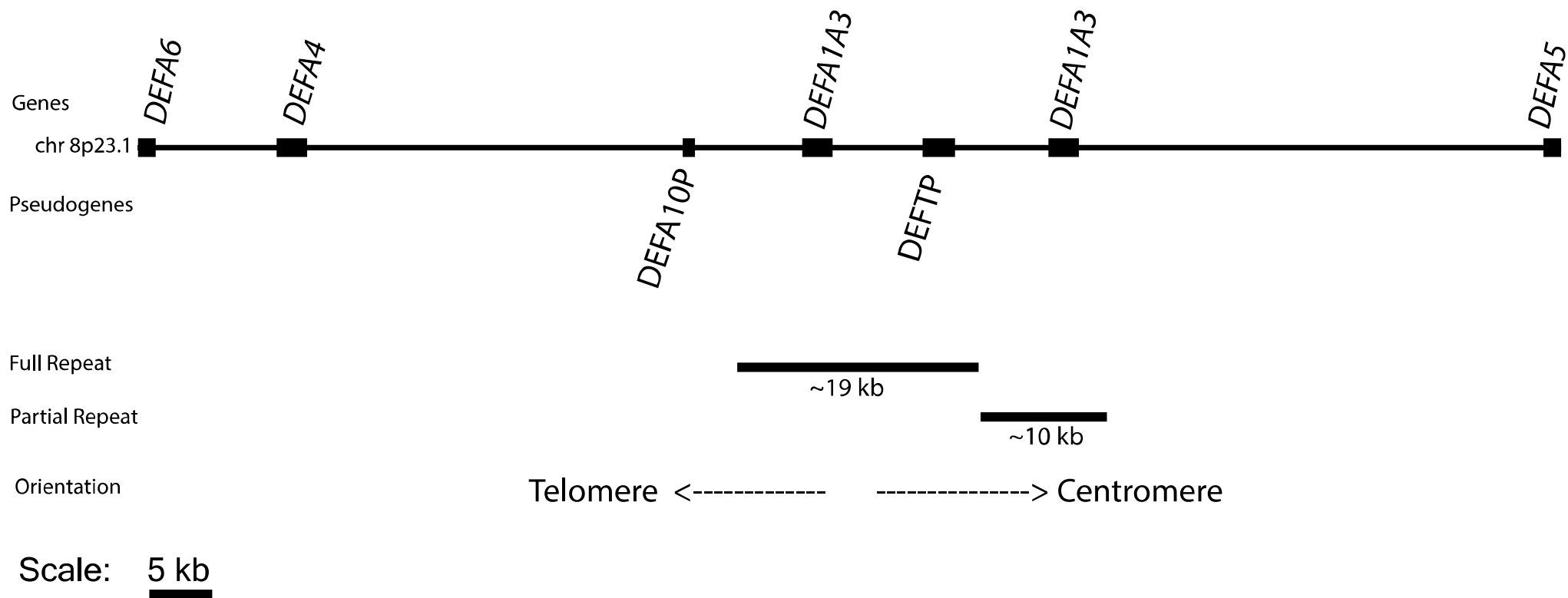
The neutrophil defensins have been shown to facilitate phagocytosis, induce histamine release and inhibit the complement pathway (Befus *et al.*, 1999, Groeneveld *et al.*, 2007, Soehnlein *et al.*, 2008). They have also been shown to be anti-inflammatory when released by apoptotic neutrophils (Miles *et al.*, 2009). They exert this effect by inhibiting the release of pro-inflammatory cytokines by macrophages even in the presence of bacteria and without killing the macrophages. Of course these activities are demonstrated *in vitro* or at best in murine models and thus how significant they are *in vivo* remains questionable. Moreover, mice do not express alpha defensins in their neutrophils. However, a study has been done looking at mice knockouts for Defb1, a beta defensin that is expressed normally in murine lungs (Morrison *et al.*, 2002). No overt deleterious phenotype in these mice was observed except that they harboured a greater load of *Staphylococcus* bacteria in their urinary bladder than control mice.

DEFA1-3 have been shown to inhibit the differentiation of monocytes into macrophages, and this property may be responsible for part of the phenotype in Chronic Myeloid Leukemia where large amounts of these peptides are released by dysplastic granulocytes (Droin *et al.*, 2010).

## **1.4. Human Alpha Defensin Genes and their CNV**

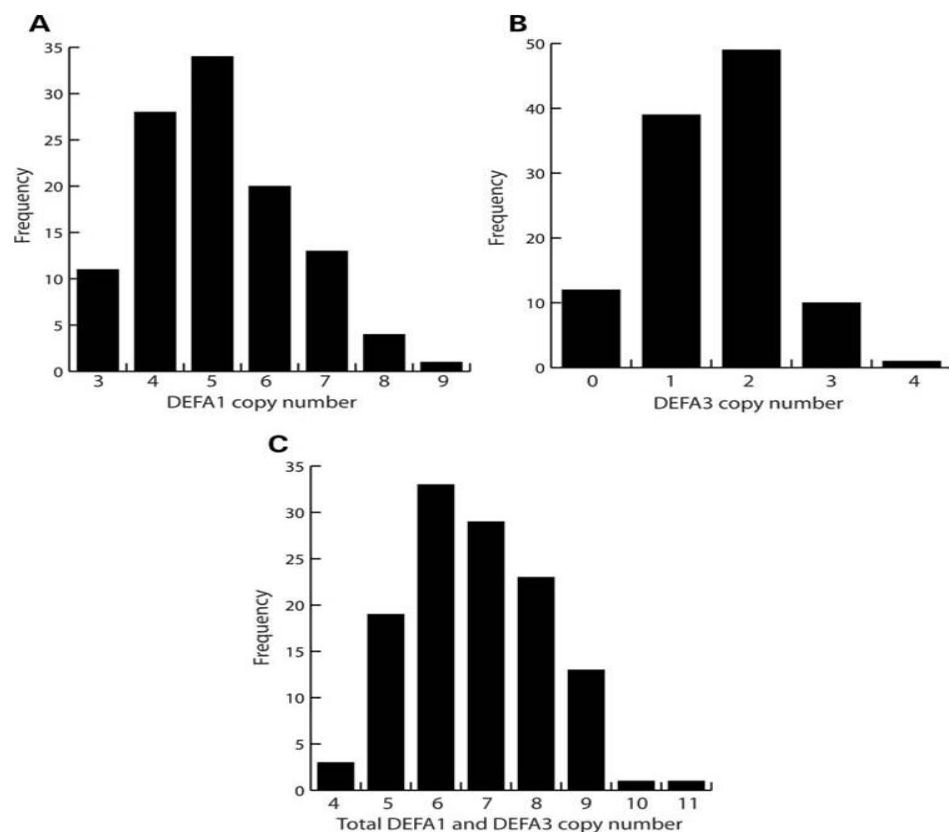
There are five known alpha defensin genes in humans, *DEFA1*, *DEFA3*, *DEFA4*, *DEFA5* and *DEFA6*, all located on chromosome 8p23.1. Since all these except *DEFA3* have syntenic orthologues in the chimpanzee genome, and since *DEFA3* varies from *DEFA1* by a single base pair change, *DEFA3* is thought to have arisen in humans via a mutation in the *DEFA1* gene. This single base pair change (C to A) results in changing the first amino acid in the mature peptide from Alanine (GCC) in *DEFA1* to Aspartic Acid (GAC) in *DEFA3*. Moreover, these two genes share a common locus such that any given copy can either

have *DEFA1* or *DEFA3*. This locus has thus been termed *DEFA1A3* and is part of a 19kb copy number variable block (Aldred *et al.*, 2005). This tandemly repeated 19kb DNA block also includes a theta defensin pseudogene (**Figure 1.6**). A 10kb partial version of this block, not including the pseudogene, is present in one copy per chromosome 8 from observations so far. There is only one reported partial deletion of this locus on dbSNP catalogued as rs71776817; however there is no information on its frequency or validation. Thus the copy number of the *DEFA1A3* locus is taken as the combined copy number of the 19kb repeat block (termed 'full repeat') plus the 10kb partial repeat.



**Figure 1.6** The Human alpha defensin gene cluster on chromosome 8p23.1 drawn to scale. The 19 kb full repeat is copy number variable with numbers ranging from 0 to 6 in European haplotypes. The 10 kb partial repeat is not known to vary in copy number. Both contain a copy of *DEFA1A3*.

The *DEFA1A3* CNV was discovered while making an attempt to characterize the genes coding for *DEFA1-3* and to find whether *DEFA1* and *DEFA3* are products of alleles or separate genes (Mars *et al.*, 1995). The copy number variable status of this locus complicates this definition. MAPH-based measurements of *DEFA1A3* in 111 European individuals using 7 calibrators typed with restriction digestion, pulsed-field gel electrophoresis and Southern blotting has shown a copy number range from 4 to 11 copies (Aldred *et al.*, 2005). From this same study *DEFA3* has been found to vary from 0 to 4 copies; 10% of the people tested did not have the *DEFA3* gene.



**Figure 1.7** Copy numbers of *DEFA1*, *DEFA3* and total (*DEFA1+DEFA3*) and their frequencies in 111 Europeans as measured by MAPH (figure taken from Aldred *et al.* 2005)

The percentage of individuals without *DEFA3* is between 10% and 15% in European and Asian (Chinese/Japanese) populations and about 37% in the African population (Ballana *et al.*, 2007). This means that any given chromosome 8 can have from 1 to 6 copies of *DEFA1* and from 0 to 2 (or maybe 3) copies of *DEFA3*. Their arrangement in the array of 19kb repeats remains elusive. *DEFA3* may have a propensity to be present at one or the other end of the array, or due to recombination and gene conversion events it

could be completely randomly situated. This idea of repeat array structure was investigated by Aldred et al. using long PCR to assay the end repeat positions for *DEFA1A3*. It was found that although *DEFA3* was more likely to be at the 5' end (centromeric end) of the repeat array, it was found within the array and not on an end repeat more than half the time. They also investigated the copy number for *DEFA1* in other primates using MAPH and found high copy numbers like 8 in chimpanzee, 11 in orang-utan, 3 in gorilla and 5 in gibbon. They only typed a single individual of each species. From array CGH experiments, CNV of the chimpanzee chromosome 8 region orthologous to the alpha defensin cluster of humans has been observed (Perry, Yang *et al.*, 2008). This implies that the alpha defensins have been copy number variable since before the divergence of humans and chimpanzees, and possibly much earlier. Twelve functional alpha-defensin genes in the marmoset sequenced genome have been identified from comparative analysis with human and other primate sequences (Das *et al.*, 2010). Marmoset is a New World Monkey (platyrrhine) and is thus a representative of this group of primates. If these 12 genes are indeed functional, then they represent either 12 unique alpha-defensin genes or include copy number variable genes, like the *DEFA1* gene in humans and chimpanzees.

For the rest of the alpha defensin genes in humans, there is no known copy number variation except for a deletion polymorphism of *DEFA4*. This is a 7 kb deletion encompassing the entire gene and is catalogued as rs71926333 on dbSNP. This deletion has resulted from the non-allelic cross-over between a 300 bp sequence on its centromeric end and a duplication of this sequence on the telomeric end. As the results from this work have shown (see Section 2.10), this is a rare deletion.

## **1.5. Alpha Defensins and Disease: Effect of CNV**

Due to their anti-infective and immunomodulatory properties, the defensins have been hypothesized and investigated as possible contributors to various diseases especially those of an autoimmune nature. These are diseases with several but unknown contributing factors. In the case of psoriasis,  $\beta$ -defensin has been shown to have a role as cited earlier. Such studies either look for correlation between peptide levels and disease, or between gene copy number polymorphism and disease.

A study looked at levels of neutrophil  $\alpha$ -defensin and a beta defensin in plasma of systemic lupus erythematosus (SLE) patients as compared to healthy controls (Stoeger *et al.*, 2009). They found significantly higher alpha-

defensins in patients versus controls, and also found a link with severity of disease. No significant correlation was observed for the beta defensin. SLE is an autoimmune disease that affects various tissues and organs in the body but with varying symptoms. The underlying cause is not known. Whether high defensin levels contribute to disease severity or are just an effect remains unanswered. If higher defensin levels do have a link to disease severity, it might be associated with higher genomic copy numbers which would give more confidence in the association. Another possibility could be that defensin levels do matter but are more a function of neutrophil count or defensin induction than genomic copy number.

Since defensins have been shown to have reduced antimicrobial activity at high salt concentrations, their role in cystic fibrosis has been hypothesized and in the case of beta defensins, has been studied. Cystic fibrosis (CF) is a monogenic, autosomal recessive disease but shows great variability in phenotype. The respiratory tract is the mainly affected system, resulting in recurrent and chronic infections, inflammation, clogging of airways and fibrosis of tissue. However the severity of lung disease is the most variable phenotype even within patients harbouring the same mutation (Davis, 2006). This points to the existence of other genetic factors that modify disease outcome. Beta defensins are expressed in the epithelial cells of the airways and high salt concentrations of CF patients' mucus would inactivate them. No association between beta defensin CNV and CF disease severity was observed (Hollox *et al.*, 2005).

There has been considerable interest in the link between defensins in the gastrointestinal tract and inflammatory bowel diseases (IBDs). Antimicrobial peptides are accepted to play an important role in maintaining the normal gut flora. While DEFA5 and 6 are normally produced in the intestinal epithelial cells, expression of DEFA1-4 has been observed in active IBDs like Ulcerative Colitis and Crohn's disease (Cunliffe, 2003). Neutrophil migration into intestinal tissue during inflammation is also a source of these defensins. On the other hand, reduced DEFA5 expression has been linked with Crohn's disease (Ramasundara *et al.*, 2009). Despite these studies nothing conclusive can be said of their role in IBDs although some role seems likely.

There has been no study attempting to associate *DEFA1A3* CNV with these diseases, the reason most likely being absence of reliable typing methods. From conflicting findings on beta defensin CNV and Crohn's, where one study associated lower (Fellermann *et al.*, 2006) and another higher (Bentley *et al.*, 2010) CNs with disease, it is quite clear that inaccurate typing can lead to false



associations. A thorough study using PRT has shown no significant association between beta defensin CNV and Crohn's disease (Aldhous *et al.*, 2010).

The WTCCC CNV study is a large-scale study of genome-wide CNVs using array-CGH in thousands of cases of 8 complex diseases, including Crohn's disease and rheumatoid arthritis (Craddock *et al.*, 2010). However, the list of CNVs they successfully genotyped does not include the *DEFA1A3* CNV. However, they genotyped the beta defensin CNV and from well-powered analysis showed that it has no association with Crohn's disease. Essentially, the role of the *DEFA1A3* CNV in diseases, if any, remains unexplored, and it even remains unknown how gene copy number polymorphism affects alpha defensin protein levels. There is one published study looking at neutrophil defensin amount and *DEFA1A3* copy number, and shows eight such measurements with a general increase in protein amount with increasing copy number (Linzmeier *et al.*, 2005). As mentioned earlier there are reasons to not have high confidence in the copy numbers measured by real-time PCR for highly variable loci such as *DEFA1A3* as has been done by Linzmeier *et al.*

## **1.6. Aims of This Thesis**

In this study, the primary aim was to develop a multi-assay PRT-based measuring system for assaying the multiallelic *DEFA1A3* CNV in the European population. Due to the non-availability of extensive genomic sequence variation information for human populations other than the Europeans at the time, it was of secondary interest to see how well this system worked in the HapMap Asian (Chinese and Japanese) and African (Yoruban) samples.

The second aim was to study the haplotypes of this locus in Europeans and obtain an estimate for the mutation rate of this CNV. Segregation information of this locus in three-generation CEPH families would allow an estimate of its mutation rate and definition of existing European haplotypes in terms of copy number. Copy number haplotypes of first generation individuals in these families can be defined further by combining them with their SNP genotypes in the same LD block (from the HapMap database).

A third aim was to test for potential functional consequences of this CNV. Given that the primary role of these peptides is in defence against infections due to their antimicrobial nature, a system for studying the effects of this CNV on resistance to infection is needed. Cystic Fibrosis (CF) patients provide such a system where all patients, despite the same CF-causing mutation, present with a highly variable infectious state phenotype. Thus it was aimed to test whether copy number of this locus associates with lung disease in CF.

Finally, it was aimed to develop a protocol for testing the relationship between gene copy number and expressed protein levels in neutrophils.

## 2. METHODS AND MATERIALS

### 2.1. Copy Number Measurement Assays

#### PRT MLT1A0

This PRT makes use of an LTR sequence in the Full Repeat unit, which is dispersed across the genome and thus provides paralogous sequences to be used as reference. Primers have been designed that amplify specifically across the MLT1A0 copy in the alpha defensin repeats and across a similar sequence on chromosome 1 (Figure 2.1). The PCR products from the two sites differ in length by 3 base pairs which allows their separation via capillary electrophoresis. This PRT excludes the Partial Repeat unit and thus we assume that the copy number measured is two less than the total copy number of *DEFA1A3*.

The forward primer has been used labelled with two fluorescent dyes, FAM or NED. Two separate PCRs are carried out for each sample; one employs the forward primer labelled with FAM and the other with NED. This gives two measurements for each sample, the average of which is finally used in the analysis. Using two separate dyes means that both PCR products for each sample can be analyzed in a single electrophoretic capillary, instead of performing the PCR twice for a single dye and then using two separate electrophoretic lanes for their analysis.

The 5' to 3' primer sequences are as follows:

MLT1A0-F: FAM/NED-CCCAGAGAGCTCCTTC

MLT1A0-R: GTGACTTATAAACAACAAAAA

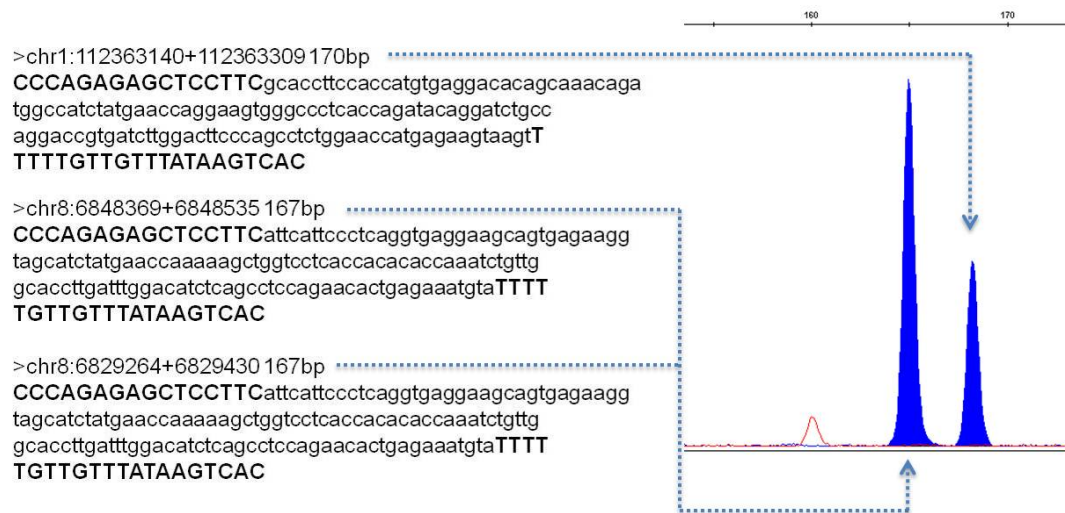
PCR:

23 cycles of:

1. 95°C for 30 seconds
2. 51°C for 30 seconds
3. 70°C for 30 seconds

1 cycle of:

4. 70°C for 30 minutes



**Figure 2.1** Predicted PCR products from PRT MLT1A0 primers and their corresponding peaks obtained after capillary electrophoresis. The genome assembly used has two full and one partial repeat, and these primers amplify only from the full repeats and the reference sequence on chromosome 1.

## PRT DEFA4-406

This PRT has been developed employing the sequence similarity between the *DEFA1A3* genes and the paralogous *DEFA4* gene. Unlike PRT MLT1A0, this PRT measures all copies of the gene because the primers amplify across the gene itself. However, the similarity in sequence also meant that designing a PRT in which the products sufficiently differed in size was not possible. The products from the two loci differ by only 2 bp (404 and 406 bp). These are not separated sufficiently on electrophoresis (Figure 2.2), and therefore a restriction digestion step has been included that uses the enzyme *MspI*, which cuts both PCR products but at different positions resulting in 275 and 317 bp products.

Only a single dye has been used for this PRT (HEX).

The 5' to 3' primer sequences are as follows:

DEFA4-406-F: TGCTCCTGCTCTCCCTCCT

DEFA4-406-R: HEX-TTGAATCAAGTCTTTGGAGAAA

PCR:

26 cycles of:

1. 95°C for 30 seconds

2. 56.5°C for 30 seconds
3. 70°C for 30 seconds

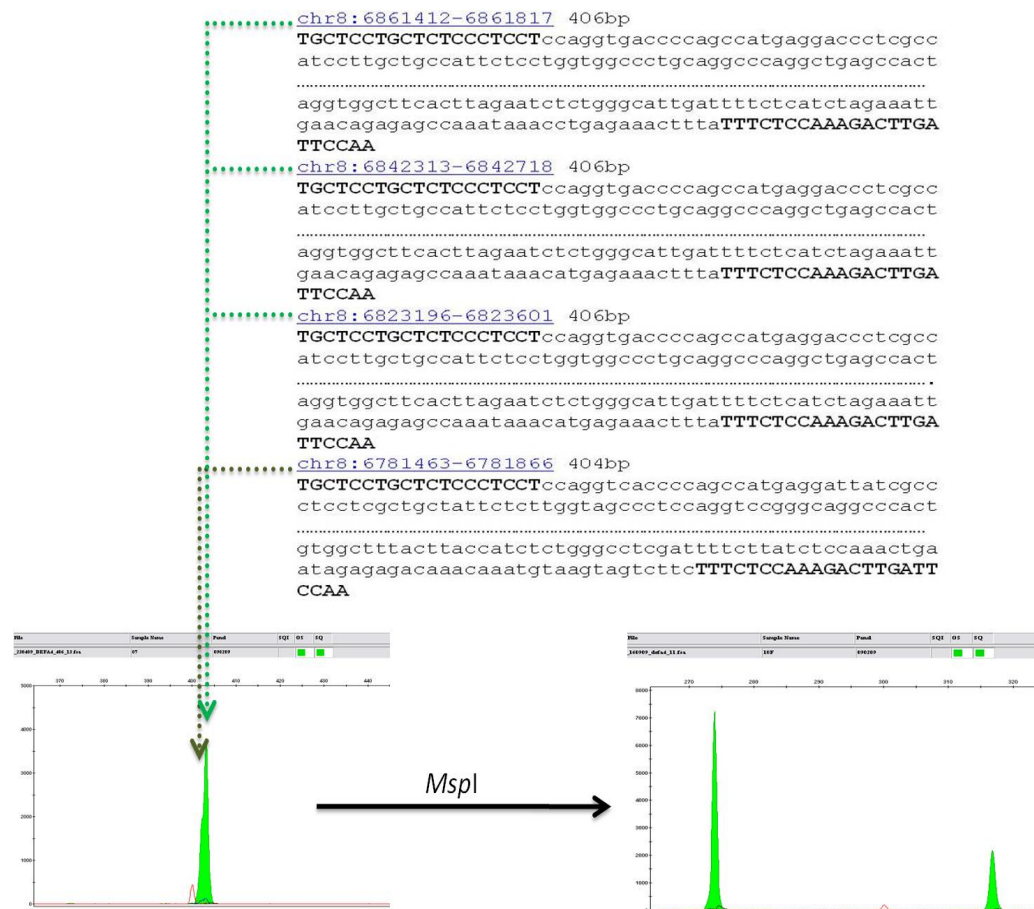
1 cycle of:

4. 70°C for 30 minutes

*MspI* digestion:

5µl PCR product + 10µl buffer/enzyme mix (3 Units enzyme per reaction)

Incubate overnight at 37°C.



**Figure 2.2** Predicted PCR products from PRT DEFA4-406 primers, their corresponding overlapping peaks after electrophoresis and well separated peaks after digestion with *MspI* and electrophoresis.

## Calibrating the PRTs

A set of 7 samples has been used to calibrate the ratios obtained from both PRTs and convert the ratios to copy numbers. These samples range in copy number from 5 to 9 which were ascertained using restriction digestion, pulsed

field electrophoresis and Southern blotting (Aldred *et al.*, 2005). Thus these samples were included in every PRT test performed.

## **2.2. Copy Number Validating Assays (Allele Ratio Assays)**

### **Indel-5 Assay**

A 5 bp insertion/deletion polymorphism in the alpha defensin repeats was identified from the NCBI trace archive sequences. The ratio of sequences containing the 5bp to those not containing it in the archived traces was about 1 to 2.5. Primers were designed to make a product across this polymorphism of size 124 bp from the shorter repeats and 129 bp from the longer ones. This assay includes both the full and partial repeats.

The forward primer has been labelled using two dyes, HEX or NED.

The primers are as follows:

Indel-5-F: HEX/NED-CTGTCCAGGAAGAGGGAGAG

Indel-5-R: CAGCTGGAGGGTCTCTGTTC

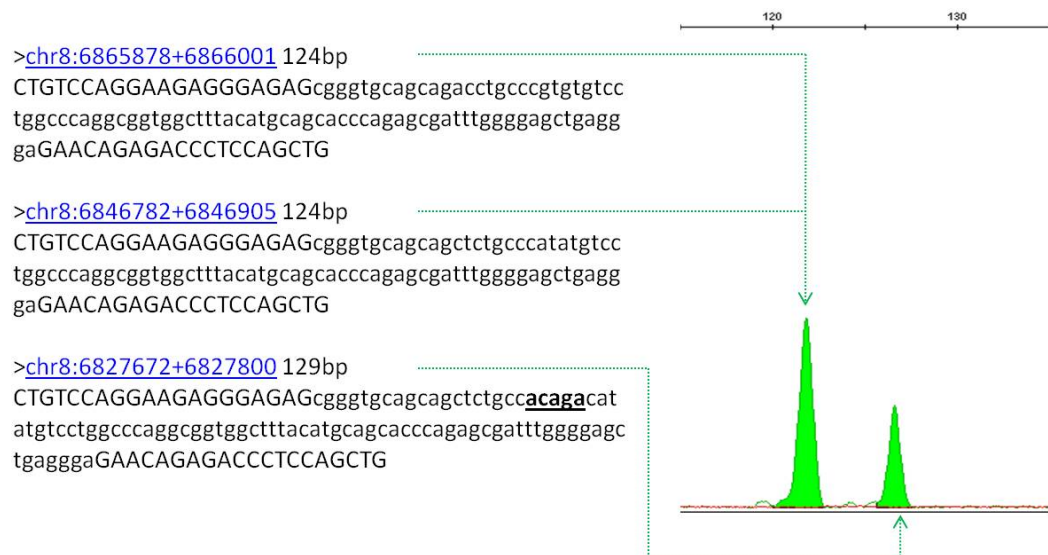
PCR:

24 cycles of

1. 95°C for 30 seconds
2. 57°C for 30 seconds
3. 70°C for 30 seconds

And 1 cycle of

4. 70°C for 30 minutes



**Figure 2.3** Predicted PCR products from Indel-5 primers and actual peaks obtained from capillary electrophoresis. The 5bp Indel sequence, in bold and underlined, is present in only one of the repeats on the genome assembly.

## DefHae3 Assay

This assay has been described elsewhere (Aldred *et al.*, 2005). The primers used amplify across the single base pair that differs between the *DEFA1* and *DEFA3* genes and alters the first codon in the mature peptide's coding sequence from GCC to GAC so that the N-terminal Alanine in *DEFA1* changes to Aspartic Acid in *DEFA3*. This base is part of a *HaeIII* recognition site in *DEFA1*, which is lost in *DEFA3*. Digestion of PCR products with this enzyme results in two different sized products. One of the primers has an introduced mismatch to create a constant *HaeIII* recognition site to act as a control for completeness of digestion (underlined base below). The reverse primer is tagged with the fluorescent dye FAM.

DefHae3-F: TGTCCCAGGCCCAAGGAAAA

DefHae3-R: FAM-TCCCTGTAGCTCTCAAAGCA

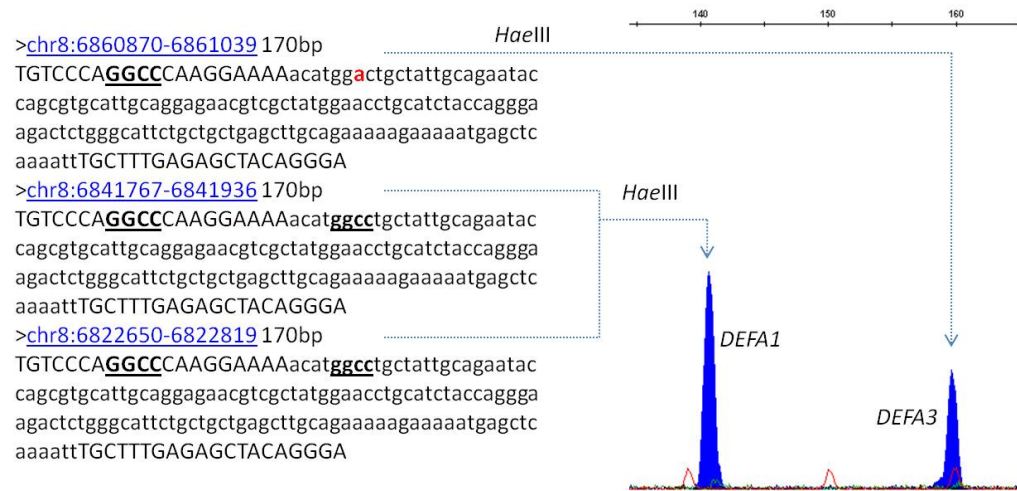
## PCR

24 cycles of:

1. 95°C for 1 minute
2. 58°C for 1 minute
3. 72°C for 1 minute

## HaeIII digestion

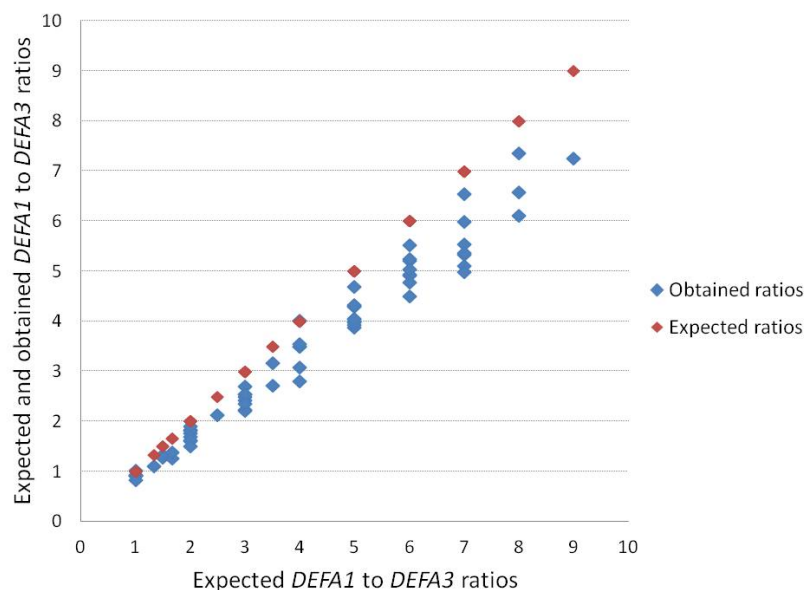
5µl PCR product + 10µl enzyme/buffer mix (0.75 Units enzyme per reaction)



**Figure 2.4** Predicted PCR products from DefHae3 primers and actual peaks obtained from capillary electrophoresis after digestion of PCR products with *HaeIII*. The 4bp *HaeIII* recognition sequences, in bold and underlined (GGCC), are present in the primer sequence as a check for digestion and also present in the internal sequence in case of *DEFA1* genes but destroyed in *DEFA3* genes due to the base substitution of C to A (highlighted in red).

Since only a single base differs between the two products of this assay, they have the potential to form heteroduplexes late in the PCR. These are then not cleaved at the distinguishing site by *HaeIII* in the digestion reaction. When made single stranded again for electrophoresis, all molecules of heteroduplexes, both *DEFA1* and *DEFA3* products, are thus the size expected from *DEFA3*. This causes a shift in the ratios obtained from the peaks towards more *DEFA3* than actual. When looking at ratios of *DEFA1* to *DEFA3*, the ratios are systematically lower than expected, by a factor of about 15%, presumably due to failure of heteroduplex to digest with *HaeIII* (Figure 2.5). When analyzing results, this error has to be accounted for.





**Figure 2.5** Graph showing deviation of obtained *DEFA1* to *DEFA3* ratios from expected ratios

## 2.3. Capillary Electrophoresis

The PCR products from both PRTs and both Allele Ratio assays were analyzed using the ABI 3100 Genetic Analyzer, which separated the single-stranded products via capillary electrophoresis and quantified the amounts via fluorescence detection. Three different fluorescent dyes, FAM, HEX and NED, have been used to label the primers. To prepare the samples for analysis they were added to 10µl mixture of HiDi Formamide and ROX-labelled markers (ratio of 170µl to 2µl for sixteen samples) and heated at 96°C for 3 minutes to ensure separation of DNA strands. The amount of sample added and electrophoresis variables used were different for different assays and are given below:

Assay	PCR product used (μl)	Voltage (kV)	Injection Time (sec)
PRT MLT1A0 <sup>α</sup>	0.7 each of FAM and NED	1	30
PRT DEFA4-406 <sup>β</sup>	1	1.5	30
Indel-5 <sup>α</sup>	0.5 each of HEX and NED	1	30
DefHae3 <sup>β</sup>	1.5	1.5	30

α: These were combined in later assays and run together

β: These were combined in later assays and run together

## 2.4. Assigning Copy Numbers

### Maximum Likelihood Program

The copy number assigned to each sample (Most Likely Copy Number: MLCN) is the consensus of the four assays described previously. This was done using a computer program (written by Prof. John Armour) that calculated the likelihood of all possible copy numbers (3 to 13 for European samples) for each sample, and took into consideration unrounded copy numbers measured from both PRTs, the standard deviation for each copy number measurement for each PRT and the peak areas from the allele ratio assays. The likelihood of each copy number given the population frequency is not included in the calculation for maximum likelihood (i.e., an assumption of “flat priors” is used). For each sample a minimum ratio is calculated which is the ratio of the likelihood of the most likely copy number (always 1) to the likelihood of the second most likely copy number. Thus, a high minimum ratio means higher confidence in the copy number assigned because the next most likely copy number has a low likelihood.

DefHae3 data was corrected before adding to the program. The percent deviation from expected ratios for different ratios was calculated and averaged across the same ratios. Apart from an expected ratio of 1 between *DEFA1* and *DEFA3*, the percent deviation was found to be about 0.16 or 16% for other ratios. The actual average including expected ratios of 1 came to be

at 0.155 and excluding ratios of 1 came to be 0.167. Thus all peak areas for *DEFA1* were multiplied by 0.15 and added back to the unaltered peak areas for correction, as they are on average 0.15 times less than the actual due to reasons discussed in the DefHae3 assay description.

## 2.5. DEFA1A3 del Assay

This assay is a simple PCR with primers designed to test for the presence of a 580 bp deletion in the *DEFA1A3* locus catalogued on the database as a SNP, rs71776817.

The 5' to 3' primer sequences are as follows:

DEFA1A3 del-F: AAGCTCAGCAGCAGAATGC

DEFA1A3 del-R: CTTGCATGGGACGAAAGCTT

Expected product size from the undeleted copies is 727 or 731 bp, depending upon another 4 bp polymorphism in the locus, and from the deleted version is 147 or 151 bp.

### PCR

95°C for 3 min

36 cycles of:

95°C for 30 sec

63°C for 30 sec

70°C for 30 sec

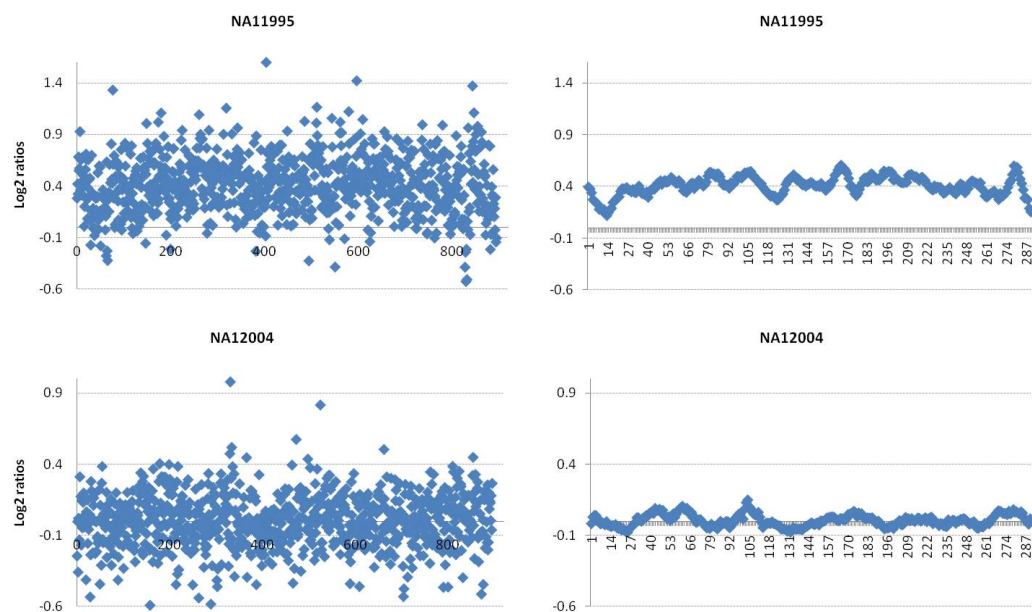
Products were analyzed by electrophoresis on an agarose gel.

## 2.6. Correlation with microarray data

A human genome-wide copy number variation study using high-resolution Comparative Genome Hybridization was undertaken by the Genome Structural Variation Consortium (Conrad *et al.*, 2010). The array used contained 42 million oligonucleotides, and 20 individuals of European ancestry and 20 of African were analyzed in comparison to a European individual. The data, binary logarithms of hybridization ratios, was made

available online after quality control and normalization (<http://www.sanger.ac.uk/cgi-bin/humgen/cnv/42mio/downloadBigDB.cgi>).

The reference individual and 18 tested individuals (9 European and 9 African) were part of the HapMap sample sets typed for *DEFA1A3* copy numbers (see Chapter 3). Copy number variation of the *DEFA1A3* containing repeat could be observed just by plotting the values for the probes covering this region on a graph (Figure 2.6). This gives a noisy picture, but if the median for each three consecutive values is taken, and then the average of ten of the medians is used, a clear picture of gain or loss is observed. For comparison of the data with the measured copy numbers the average of the  $\log_2$  ratios for all probes covering the alpha defensin repeat region could be calculated and predicted to have a correlation with measured copy numbers. Inversely, the anti-log of the averaged values of a sample could be predicted to be similar to the ratio of that sample's copy number to the reference sample's copy number (see Chapter 3).

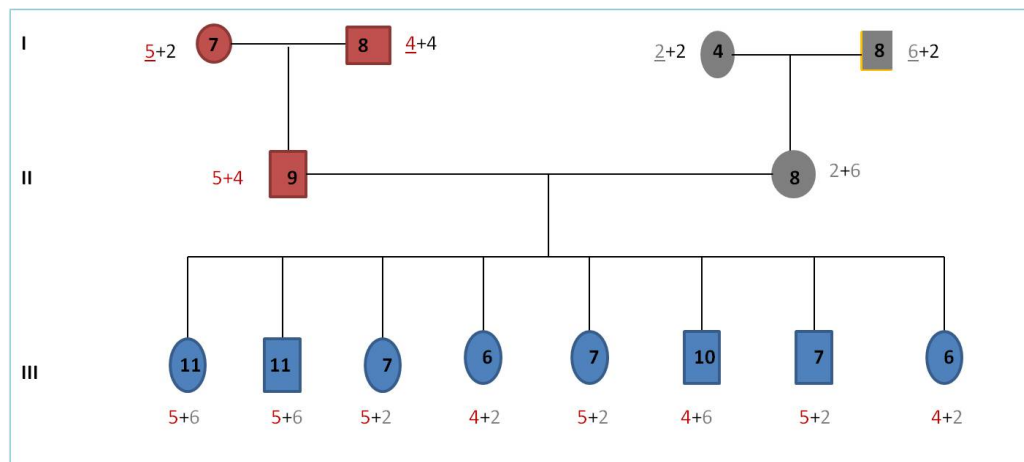


**Figure 2.6** On the left are the plots of the raw data for the probes covering the region chr8: 6,816,648-6,867,537, and on the right are the averages of the medians for the same region in the same two samples.

*(This analysis makes use of data generated by the Genome Structural Variation Consortium (PIs Nigel Carter, Matthew Hurles, Charles Lee and Stephen Scherer) whom we thank for access to their CNV discovery and genotyping data, made available through the websites <http://www.sanger.ac.uk/humgen/cnv/42mio/> and <http://projects.tcag.ca/variation/> as a resource to the community.)*

## 2.7. Segregation Analysis

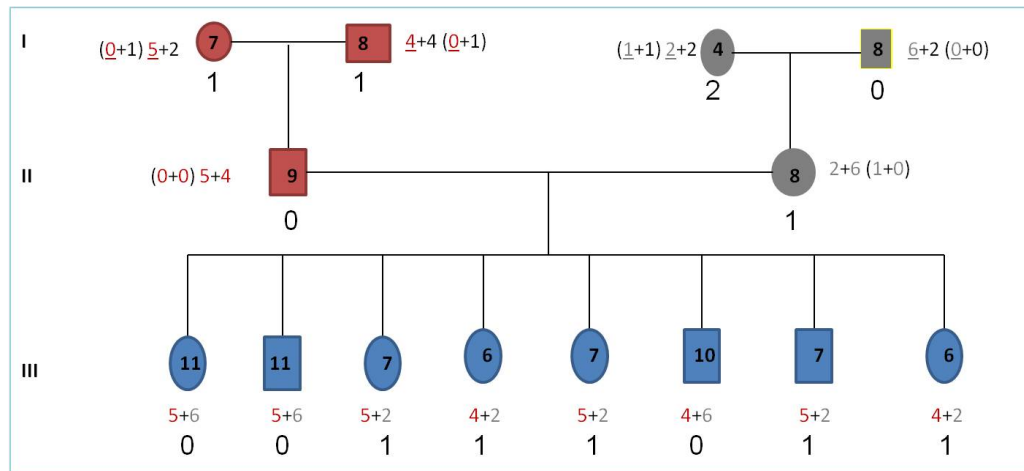
Three-generation CEPH pedigrees were typed with the four assays for the purpose of inferring haplotype copy numbers of *DEFA1A3* on the principle of segregation of chromosomes from parents to offspring. The assignment of haplotype copy numbers was done manually based on measured diploid copy numbers, diploid Indel-5 copy numbers, diploid *DEFA1/DEFA3* copy numbers and segregation information from surrounding markers. Counterintuitively, high diploid copy numbers, such as for the *DEFA1A3* locus, rarely present with a situation where more than one combination of haplotypes is possible in the parental generation given the availability of a large number of children, exhibiting, usually, all four possible combinations of haplotypes. For example, consider diploid copy numbers of 5 and 3 for a hypothetical locus in a set of parents who have children with four different diploid copy numbers of 2, 3, 5 and 6. In this scenario, two sets of haplotypes in the parents can be possible: 2+3 and 0+3, and 4+1 and 2+1. As diploid copy numbers increase (e.g., to a range such as that exhibited by *DEFA1A3*), possible haplotype combinations decrease to usually one possibility. An example is presented in Figure 2.7.



**Figure 2.7** CEPH Family 1408. Measured diploid copy numbers within symbols and haplotype copy numbers outside. Black shows non-transmitted haplotypes in generation I. Red haplotypes in generation III inherited from father and grey haplotypes from mother.

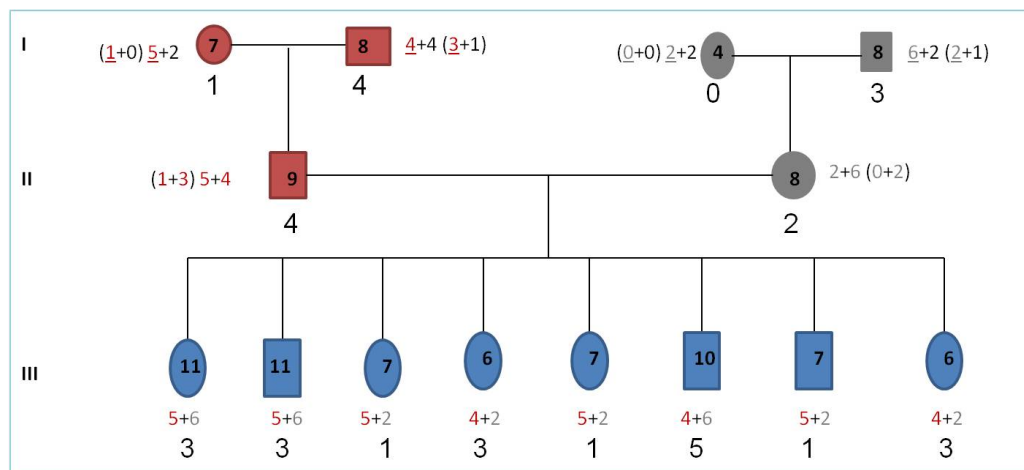
In Figure 2.7 *DEFA3* and Indel-5 alleles' copy numbers are not shown, although they are also useful for haplotype sorting and their own haplotype copy numbers can also be inferred in a similar manner. As shown in Figure 2.8, the *DEFA3* (and hence also the *DEFA1*) haplotype copy numbers were easily assigned based on the measured diploid copy numbers of these genes from the DefHae3 assay. The *DEFA3* diploid copy numbers were arrived at given the

MLCN for each sample, e.g. a *DEFA1* to *DEFA3* ratio of 1 for a sample with total copy number of 4 results in assigning 2 *DEFA3* copies to the sample (as is true for the maternal grandmother in the figure below, who has a total copy number of 4 and a *DEFA3* diploid copy number of 2).



**Figure 2.8** CEPH Family 1408. The diploid *DEFA3* copy numbers are shown below the symbols in black. The haplotype *DEFA3* copy numbers are given in brackets next to the total haplotype copy numbers in generation I and II.

Similarly the Indel-5 alleles' copy number from the Indel-5 allele ratio assay could be put to use. For the same family, family 1408, Figure 2.9 shows the Indel-5 copy number assignment. It is a little more complicated and relies more on the copy numbers found in the third generation, much like the assignment of total haplotype copy number.



**Figure 2.9** CEPH Family 1408. The diploid copy numbers of Indel-5 minor allele (repeats containing the 5-bp sequences) are shown below the symbols in black. The haplotype copy numbers for this allele are given in brackets next to the total haplotype copy numbers in generation I and II.

In family 1408, despite the high copy numbers it was easy to work out the haplotypes mainly because of the variety of haplotypes they carry. In families where parents share haplotype copy numbers, it becomes impossible to say from segregation data only which haplotype carries the lone *DEFA3* inherited by the children when each parent has one copy, for example. In these cases, further information could be obtained using linkage analysis and known segregation patterns of linked markers.

## **2.8. SNP Association with Copy Number Haplotypes**

Some of the father-mother-child trios making up the first and second generation of the CEPH families typed were part of the Phase I CEU HapMap samples for whom phased SNP genotypes are available online ([www.hapmap.ncbi.nlm.nih.gov](http://www.hapmap.ncbi.nlm.nih.gov)). Thus 14 such trios from 9 families were available. The linkage disequilibrium block containing the *DEFA1A3* repeats extends from chromosome 8 coordinate 6,793,624 to coordinate 6,908,814 (UCSC Genome Browser assembly hg18). Excluding the children from the trios, the haplotype copy numbers of *DEFA1A3*, Indel-5 minor allele and *DEFA3* were combined with their SNP haplotypes. These were then sorted according to the copy number of each of the three measures to look for SNP haplotypes that tagged copy number lineages. Association of SNP haplotype with copy number was evaluated using a chi-squared test, where the copy number haplotypes were divided into two categories of various combinations. For example, haplotypes that had no *DEFA3* genes were compared against those that had at least one, or haplotypes with 2 or 3 copies of *DEFA1A3* were compared against those with 4, 5 or 6, and the frequency of the major allele of the SNP was calculated in both categories. Starting with the hypothesis that there would exist no difference in the SNP major allele frequencies between the different copy number lineages, those SNPs that gave the lowest p-value in each analysis were noted. Since the idea was then to test how well those SNPs replicated this copy number-tagging ability in an independent set of samples, the p-values were not corrected for multiple testing.

## **2.9. rs4300027 Genotyping Assay**

SNP rs4300027 was found to give a low p-value for a chi-squared test of independence between the major allele frequency in low-copy and high-copy *DEFA1A3* haplotypes. To test this association in an independent set of

samples, namely the ECACC samples which had already been typed for copy number, a SNP typing assay was devised. The SNP is a T to C transition. In the C allele context, it makes the recognition site GAWTC for the restriction enzyme *TfiI*. In the T allele context, this recognition site is lost. Primers were designed around this SNP to include it and another non-variable *TfiI* site.

rs4300027-F: AGATACCATGCTTGGAGGAA

rs4300027-R: GGGTCTTGAATTCAAATGTCAG

## PCR

36 cycles of:

95°C for 30 seconds

58.6°C for 30 seconds

70°C for 30 seconds

These primers give a product 1043 bp long. Digestion with *TfiI* gives 834 and 200 bp products in sequences with allele T, and 660, 200 and 174 bp products in sequences with allele C.

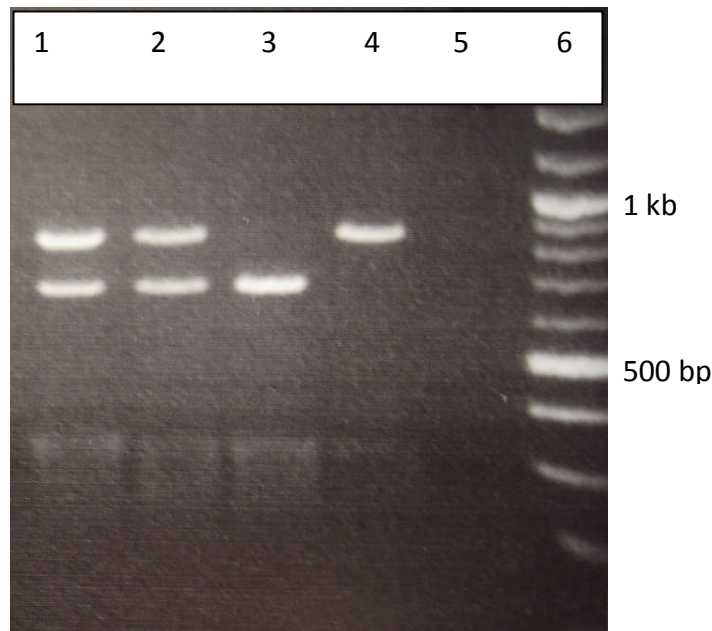
chr8:6867808+6868850 1043bp

AGATACCATGCTTGGAGGAAgactaagcatccacagggagaggaactgagcc  
cacctgcaaggatgggctagagaacactgagcaaccagctttctaggaaaaaa  
gaaaactctgatttgcaatgtttgttaaatttctgtggttaaaatgctcccagc  
tatagacagttta**gaatc**atcacacaaaaactcctccctcatgagctggcct  
gatctgaccccagcacatcacagggtctcatccttcagctttctcagagtttc  
cagctgagccaacaccacctgcccacctgtgcacgagtgctcctggccctgaaat  
tttcagatctcagcagaacctctcctcttatgcccgtggaaggatccaaacc  
caattgcaaagtgtgtgagtgaagacgtgatcatgctgtttcaatccactactt  
tctgtggtgtcttttcgcacagtcctagatgaacagaaggcaggtcttggtg  
agaagttgaatgtgtgcattttttgtgtgtgttaaattctcagcctctctataat  
attgttgaagtaggacagaacctctcaccttattttccaaagtgtcacaaaga  
gcccattctaatggcagcgtggaattgtggactctttggagtgactgaagaac  
ccccgtcaccattcttagtttaaattcttcctgttcagagcaggggtggtgt  
gggagccaggtggagtgtaacctctccccacagtgcagagactcagaggaggc  
cacgggacttgggggttggtggaacaacatgggaagaagtagggattttctc  
caggagattagctacaaaagtcataagagagatgacgat**gattc**acatggttgt  
cacctggggagagagcccaaagtaagttcaagagaaacaaagggaaaaaaaag  
gtgaagacaacagtgttaactcatgctttgatttcttgatcaagcaattcctg  
aagctagaattcctcctcaattccaagatataggagtcaaaatgttaacttag  
cttggtgttggtttCTGACATTTGAATTCAGACCC

rs4300027 is highlighted in green in the predicted PCR product sequence above. The two *TfiI* sites are in bold and underlined. The first of these is thus destroyed in sequences which have a T at this SNP position. The second of



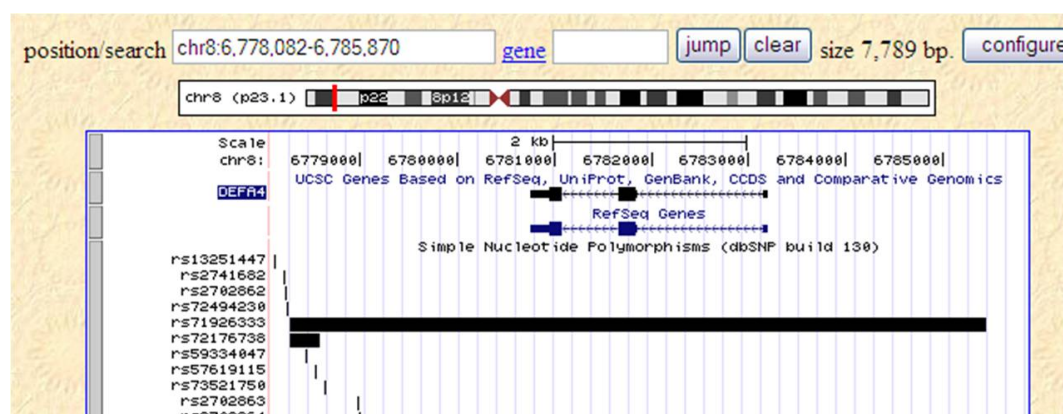
these, however, remains unchanged, and thus serves as a positive check of enzyme digestion. The products from digestion were analyzed on a 1% agarose gel via electrophoresis (Figure 2.10).



**Figure 2.10** Agarose gel electrophoresis of *TfiI* digested products from the rs4300027 assay. Lanes 1 and 2 have T/C heterozygotes, lane 3 has CC homozygote, lane 4 has TT homozygote, lane 5 is a PCR negative control and lane 6 is a DNA ladder. The restriction digestion buffer has affected the electrophoretic mobility of the products, so they migrate at a slower rate than the corresponding DNA markers.

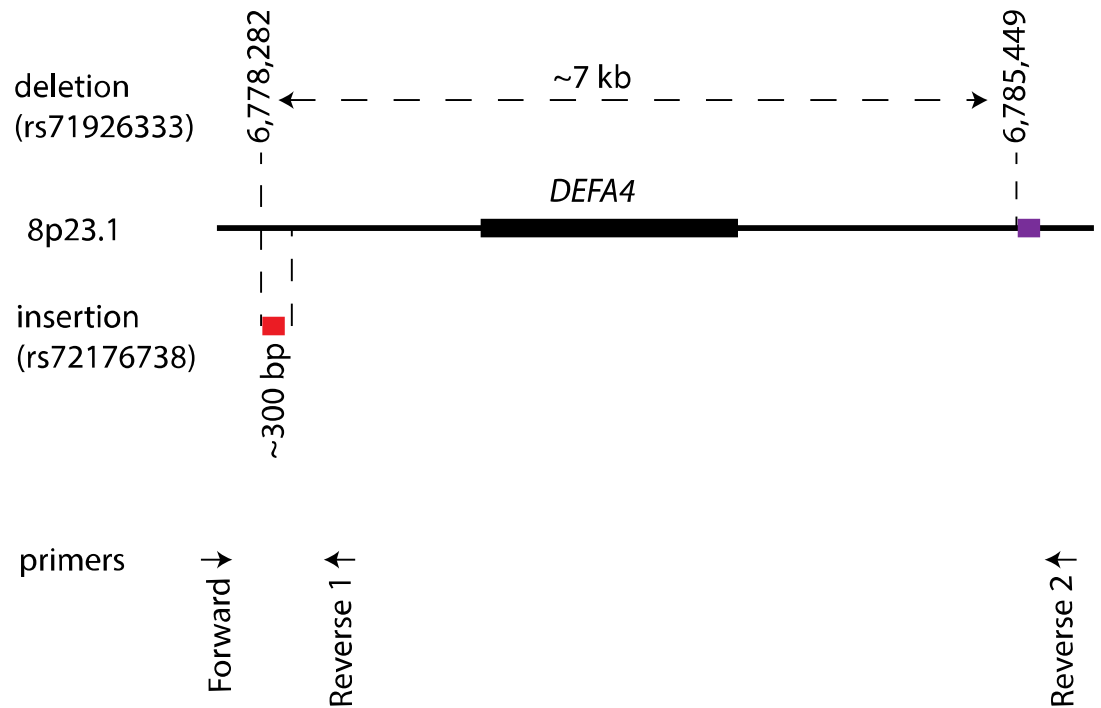
## 2.10. DEFA4 Deletion Assay

From the PRT results for CEPH family 1346, it was observed for the maternal grandfather, mother and four children that their inferred DEFA4-406 PRT copy number was twice that measured from PRT MLT1A0. One simple explanation of this discrepancy would be that they had only one copy of the DEFA4-406 PRT reference locus, the *DEFA4* gene. The UCSC genome browser shows a 7169 bp deletion polymorphism (rs71926333) covering the *DEFA4* gene, and an overlapping 302 bp deletion polymorphism (rs72176738) at the telomeric end in dbSNP build 130 (Figure 2.11).



**Figure 2.11** Snapshot of the UCSC genome browser showing the *DEFA4* gene and the 7169 bp and 302 bp deletions, both represented as SNPs.

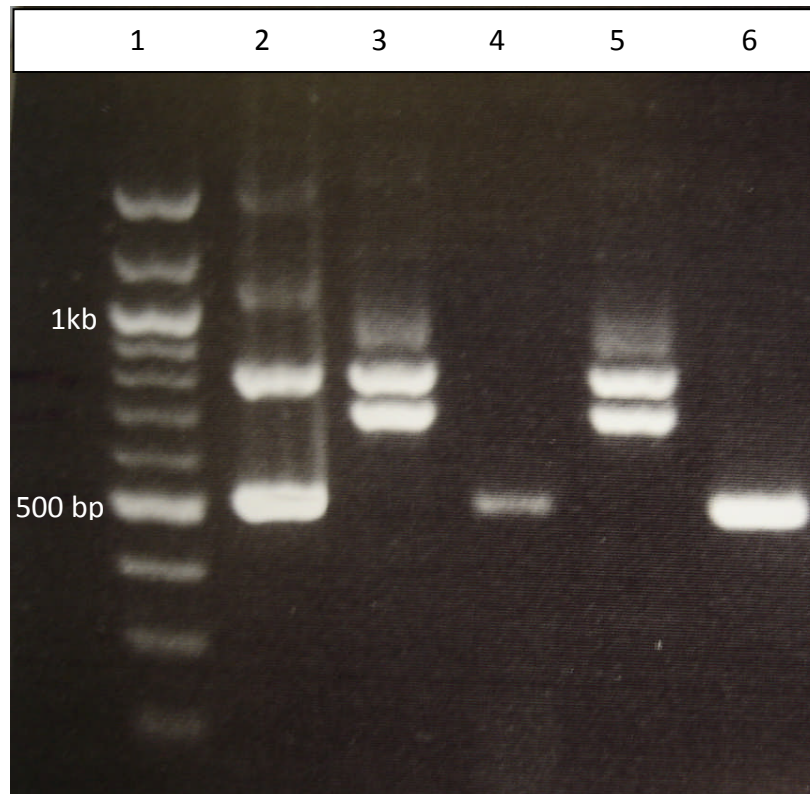
On searching the NCBI trace archives for sequences with the large and small deletions, it was found that the 302 bp deletion (rs72176738) sequence was ancestrally present in another location (purple bar in **Figure 2.12**). Also, the 302 bp deletion allele appears to be the ancestral state (so that the polymorphism might be more correctly designated as an insertion) given the frequency of its absence in the sequences on the NCBI trace archives. This explained the occurrence of the large deletion due to the non-homologous recombination between the perfectly matched 302 bp sequences at either end. A three-primer PCR assay was designed that could check for the presence of the 302 bp insertion and the 7169 bp deletion simultaneously (**Figure 2.12**). The left primer (forward primer) is located outside the polymorphic region at the telomeric side, the first right primer (reverse primer 1) within the 7169 bp sequence and the second right primer (reverse primer 2) beyond the 302 bp on the centromeric end. Three possible kinds of haplotypes can exist: one which has no insertion or deletion, a second which has the insertion, and third which has the deletion. In the first and second case, the forward and first reverse primer will give products, but the sizes will be different depending upon the presence or absence of the insertion. In case of the deletion, the forward and second reverse primer will give products (designed to be of a different size from the first two products).



Scale: — 1 kb

**Figure 2.12** Scaled representation of the deletion and insertion (duplication) polymorphism at the *DEFA4* locus. Purple bar represents the original location of the 302 bp, red bar the insertion of that sequence and the intervening sequence is deleted along with 302 bp of sequence, some from the original and some from the inserted 302 bp. Arrows denote the approximate position of primers. The Forward and Revers1 primers will give products of sizes 526 bp and 828 bp from chromosomes without and with the insertion respectively. The Forward and Reverse2 primers will give a product of size 720 bp from chromosomes with the 7 kb deletion.

Note that the forward and second reverse primer can always give a product, but in the non-deleted state (A and B in **Figure 2.12**) it is almost 8 kb in size and thus most probably not significantly amplified in the PCR nor is it in the resolving range of the 1.5% agarose gel (Figure 2.13).



**Figure 2.13** Agarose gel electrophoresis of 7kbDEL PCR. Lane 1 is a DNA marker, lane 2 is a heterozygous carrier for the insertion, lanes 3 and 5 are heterozygous carriers of the deletion and insertion, and lanes 4 and 6 are homozygous for the reference sequence.

### Primers

7kb DEL-F: CCTGATTCTACATTCTCTACCTA

7kb DEL-R1: GAATTGTAAGCTATCTCACATCT

7kb DEL-R2: CTTATGTATGTAAATGTTGTACGG

### PCR

36 cycles of:

95°C for 30 seconds

52°C for 30 seconds

70°C for 30 seconds

## 2.11. PRT DEFTP1 and usDEL Assay

Primers for DEFTP1 PRT were designed that matched a sequence in the theta-defensin pseudogene (*DEFTP1*) which is present in each full repeat, and a

paralogous sequence upstream of the first full repeat that has an annotated alpha-defensin pseudogene (*DEFA10P*) which would act as a reference (see Section 4.5). The usDEL assay was designed to genotype the apparent deletion polymorphism in this paralogous sequence, upstream of the first full repeat (see Section 4.5), which was also the reference site for PRT DEFTP1. Several primer pairs failed to give a short product from carriers of the deletion, which led to the hypothesis that it was a replacement rather than a deletion polymorphism. Primers usDEL5-F and usDEL1-R were used to amplify across this polymorphic region, and restriction enzyme digestion was used to differentiate between products from the ancestral sequence (digests with *Bst*UI) and from the replacement-carrying sequence (digests with *Eco*RI).

## Primers

*DEFTP1* PRT-F: HEX-CTGCAGCTTCATCAGCTCT

*DEFTP1* PRT-R: AAAACATGCCAGGATCCTC

usDEL5-F: TAGAGCTGGGACCATTTATG

usDEL1-R: GTTACGGTTAGAGCACAGGTAC

## PCRs

### 1. *DEFTP1* PRT

95°C for 3 minutes

26 cycles of:

95°C for 30 seconds

63°C for 30 seconds

70°C for 30 seconds

1 cycle of:

72°C for 30 minutes

### 2. usDEL5 PCR (using Phusion® Polymerase)

98°C for 30 seconds

37 cycles of:

98°C for 15 seconds

61°C for 30 seconds

72°C for 2 minutes 15 seconds

1 cycle of:

72°C for 10 minutes

### Restriction Enzyme Digestions

*EcoRI*: 5µl PCR product + 10µl Buffer and enzyme mix (1 Unit Enzyme per reaction). Incubate at 37°C overnight.

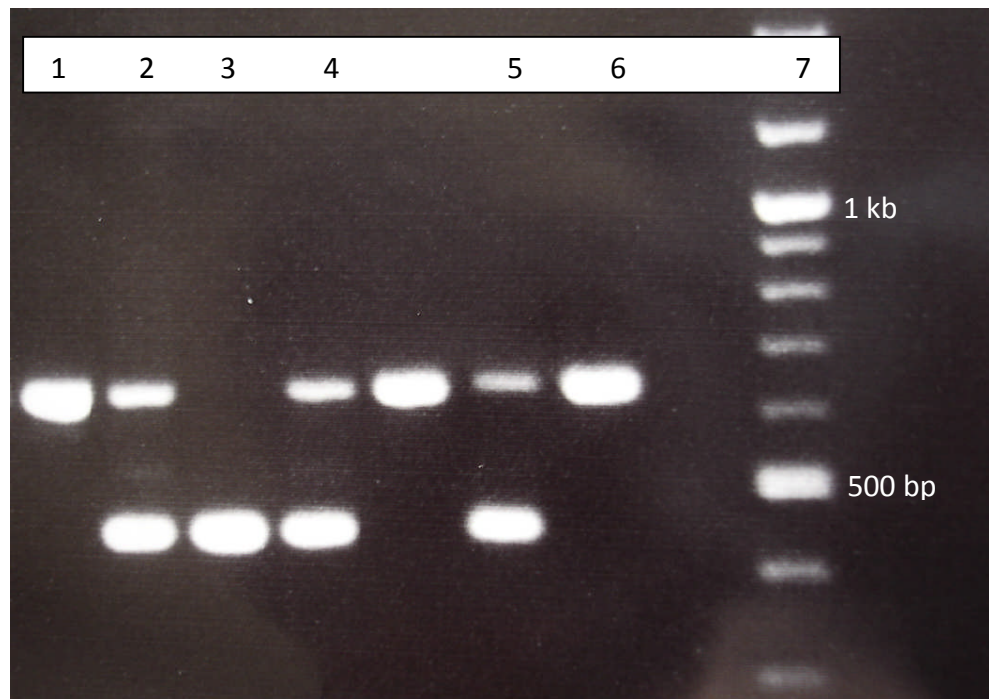
*BstUI*: 5µl PCR product + 10µl Buffer and enzyme mix (1 Unit Enzyme per reaction). Incubate at 60°C overnight.

### ABI Capillary Electrophoresis of PRT *DEFTP1*

1µl PCR product + 10µl of HiDi® and ROX marker mix. Run voltage 1kV.  
Injection time 30 seconds.

## 2.12. Upstream Sequence Replacement Assay

A three-primer assay was designed for this polymorphism (see section 4.5). The left primer (Repl1-F) was placed outside the telomeric end of the replacement junction, selected to be specific to this site and not the full repeat sequence, and two right primers were designed, one that matched the reference upstream sequence (Repl1-orig-R) and a second that matched the full repeat sequence (Repl1-repl-R).



**Figure 2.14** Agarose gel electrophoresis (1% agarose) of replacement assay PCR products. Samples in lanes 2, 4 and 6 are heterozygous for the replacement polymorphism, those in lanes 1, 5 and 7 have the reference sequence and the one in lane 3 is homozygous for the replacement.

### Primers

Repl1-F: AGCAGCAGATCCGGTATAATC

Repl1-orig-R: AGAGCCCAATAAATCTAACAGG

Repl1-repl-R: GTGAATCCAGAAAGAACGAGTC

### PCR

95°C for 3 minutes

36 cycles of:

95°C for 30 seconds

60°C for 30 seconds

70°C for 30 seconds

## 2.13. Protein Methods

### Protein Extraction

As DEFA1, 2 and 3 are present in neutrophils, blood samples served as a source for neutrophils from which proteins were extracted subsequently. The protocol was adapted from a study by Linzmeier and Ganz (Linzmeier *et al.*, 2005).

### Neutrophil Collection

Peripheral venous blood was drawn from healthy donors and collected in EDTA-containing vacutainers (BD). At least 5 ml blood was taken. Granulocytes were separated from other blood components by centrifugation, using Lympholyte-poly<sup>®</sup> by CEDARLANE<sup>®</sup> and following manufacturer's instructions. Lympholyte-poly<sup>®</sup> is a separating medium that contains dextran 500 and sodium metrizoate. Layering blood on an equal volume of this medium and then centrifuging separates blood into red blood cells, plasma, mononuclear cells and granulocytes. Of the granulocytes, neutrophils make up about 95% of cells, and eosinophils and basophils make up the rest. The neutrophils are not further purified as the presence of eosinophils and basophils does not affect downstream assays, especially since they are in small amounts as compared to neutrophils.

### Protein Extraction

The separated neutrophils were suspended in normal saline (0.9% sodium chloride). Their concentration was measured (cells per mL) using microscopy with a haemocytometer slide. Volumes containing  $5 \times 10^5$  cells were taken in 1.5 mL eppendorf tubes, centrifuged to pellet the cells and supernatants were discarded. These pellets were suspended in 100  $\mu$ l acetic acid (5%), vortexed and left shaking at 4 °C overnight. This suspension in acetic acid is the main extraction step. After this step the suspensions were dried in a centrifuge under vacuum. These dried extracts were then suspended in the appropriate buffer depending upon the assay/procedure (see below).

### Confirming Alpha Defensins in Protein Extracts

To check for the alpha defensin peptides, the neutrophil extracts were analyzed on a SDS Polyacrylamide Gel Electrophoresis (PAGE) using a tris-tricine buffer system. A discontinuous gel system consisting of a stacking gel and an 18% resolving gel, and a Mini-PROTEAN II electrophoresis cell was used. The dried extracts were suspended in protein loading buffer that contained 2% SDS, 10% glycerol, 0.01% bromophenol blue in 0.06M Tris HCl



(pH 6.8) and 0.8M DTT (Dithiothreitol). Prepared buffer pack was purchased from Fermentas for use. DTT acts as a reducing agent to disrupt disulfide bridges and bromophenol blue acts as a tracking dye during electrophoresis. The samples were heated at 100 °C for 3 minutes before loading. The electrophoresis time was about 1.75 hours, stopping when the dye had just run off from the bottom of the gel. The voltage used was 80 V initially, until the samples had entered the separating gel, and then increased to 160 V. After electrophoresis, the gel was gently rinsed in tap water to wash off buffer and then immersed in a fixing solution (40% methanol, 7% acetic acid) for one hour. Then it was stained overnight in Coomassie Blue stain. The gel was next de-stained in a de-staining solution (15% methanol, 10% acetic acid) for one hour, and then in a fresh de-staining solution for another hour.

### Gel and Buffer Recipes Used:

Gel	Acrylamide (ml)	4x Buffer (ml)	TEMED (μl)	40% Ammonium Persulphate (μl)	Water (ml)
Stacking	0.67 (A)	1 (a)	16	4	2.33
Resolving (18%)	3.6 (B)	2 (b)	12	8	2.4

A= 30% acrylamide:bisacrylamide (37.5:1)

B= 40% acrylamide:bisacrylamide (19:1)

a= 4x Stacking buffer (0.5 M Tris HCl, pH 6.8, 0.4% SDS)

b= 4x Resolving buffer (1.5 M Tris HCl, pH 8.3, 0.4% SDS)

10x Anode Buffer: 1 M Tris-HCl, pH 8.9

10x Cathode Buffer: 3 M Tris, 1 M Tricine, 1% SDS, pH ~8.25

5 μl of a low range protein ladder (from Fermentas) was also run in each gel. The expected size of the defensins band on the gel is 3.5 kDa. This band was cut out and sent for peptide sequencing, which confirmed that it contained the alpha defensins (see Section 6.3).

### Acid Urea Gel Electrophoresis

To separate the three defensin peptides (or two, in case of *DEFA3* absence) extracted from neutrophils, acid-urea gel electrophoresis was employed. The samples were same as for SDS-PAGE and ELISA (desiccated neutrophil

extracts). The sample buffer, gel running buffer and gel recipes are given below:

- **Sample Buffer**
  - Acetic acid 0.06M
  - Urea 4M
  - Glycerol 20%
  - Methyl Green 0.1%
- **Electrophoresis Buffer**
  - Acetic acid 0.9M
- **Stacking Gel**
  - Acrylamide\* 7%
  - Acetic acid 0.125M
  - Urea 4M
  - TEMED 0.5%
  - APS\*\* 0.2%
- **Separating Gel**
  - Acrylamide\* 12.5%
  - Acetic acid 0.2M
  - Urea 4M
  - TEMED 0.33%
  - APS\*\* 0.2%

\*Acrylamide used was a 37.5:1 acrylamide:bisacrylamide solution;

\*\*APS=Ammonium persulphate (freshly prepared solution)

## Pre-electrophoresis

After pouring the separating gel and allowing it to set, it was pre-electrophoresed at a constant voltage of 200V (for 18x20 cm gels) for 12 hours. The buffer was then discarded and the stacking gel was poured with the well-forming comb inserted at top.

## Electrophoresis

Any unpolymerized solution was removed from the wells. Fresh electrophoresis buffer was added. Samples were suspended in 15µl sample buffer and applied to the wells. Electrophoresis was done at 300V for 5.5 hours.

## Staining

The staining and de-staining of the gel was done following the same procedure as for SDS-PAGE described earlier.

## Alpha Defensin Peptide Quantification by ELISA

For ELISA-based quantification of DEFA1, 2 and 3 peptides, an ELISA kit was obtained from Hycult biotech<sup>®</sup> that contained all components required. It quantifies the three peptides together as the antibodies cannot discern between the single amino acid change/absence that characterizes DEFA1, 2 and 3.

### Sample Preparation for ELISA

The dried neutrophil extracts were suspended in the sample buffer provided in the kit and diluted to give results within the range covered by the standards.

### Standards and Negative Controls Preparation for ELISA

The kit included purified DEFA1, 2 and 3 mix extracted from human neutrophils in a desiccated form. This was dissolved in distilled water and 8 serial dilutions using sample buffer were made for each experiment, including one blank. A set of 8 blanks mimicking the serial dilutions for the standards were also included in each experiment. These blanks had pure distilled water instead of the dissolved standard peptides.

### Principle of ELISA

This ELISA worked on what is called the sandwich method. The standards, blanks and samples were applied to the wells which were pre-coated with the antibodies. After washing off unbound proteins, a secondary antibody to the defensin peptides was applied, thus forming a sandwich with a layer of alpha defensin peptides between layers of antibodies. These secondary antibodies had biotin attached to them. After washing off unbound antibodies, streptavidin-peroxidase conjugate was applied to the wells. After washing off unbound streptavidin-peroxidase conjugate, a substrate for the enzyme peroxidase is applied. This substrate is 3, 3', 5, 5'-tetramethylbenzidine (TMB) which is colourless. When peroxidase acts upon it, it is converted to a blue product. To stop the reaction, acid is added to the wells, which also causes the blue product to change to yellow. The absorbance of this product is measured at 450 nm. The higher the absorbance, the more biotin-conjugate antibody is present, corresponding to a larger amount of defensins in the sample applied to the well.

## 2.14. Samples, General Reagents and Procedures

### PRT Standard/Reference Samples

These samples are from the UK population extracted from whole blood using a standard phenol/chloroform method adapted from the Nucleon DNA extraction kit (Aldred *et al.*, 2005).

### ECACC Samples

ECACC (European Collection of Cell Cultures) DNA samples are random UK Caucasian control samples obtained from the Health Protection Agency ([www.hpacultures.org.uk](http://www.hpacultures.org.uk)). ECACC panel 1 and 2 were used. These DNA samples are extracted from cell lines derived by Epstein Barr Virus transformation of peripheral blood lymphocytes.

### HapMap Samples

The 90 HapMap CEPH (CEU) DNA samples are from the U.S. population with Northern and Western European ancestry, that were collected by the Centre d'Etude du Polymorphisme Humain (CEPH). The 90 HapMap Yoruban (YRI) samples come from the people of Ibadan, Nigeria. CEPH and Yoruban HapMap samples are made up of 30 trios of mother, father and child. Of the 90 HapMap Asian samples 45 come from unrelated Han Chinese from Beijing and the other 45 from unrelated Japanese from Tokyo. All HapMap DNA samples are part of the International HapMap project (Gibbs *et al.*, 2003), and have been obtained from the Coriell Institute (<http://ccr.coriell.org>). These DNA samples are extracted from cell lines derived by Epstein Barr Virus transformation of peripheral blood lymphocytes.

### CEPH Families

Seventeen three generation CEPH families of European origin were used in the segregation analysis. Five of these families had both sets of grandparents and both parents (two trios) and four of these families had just one set of grandparents and one parent (one trio) also represented in the HapMap CEPH samples described above. These DNA samples have also been obtained from the Coriell Institute (<http://ccr.coriell.org>).

## Cystic Fibrosis Samples

These DNA samples were kindly provided by Ed Hollox, and had been extracted from peripheral blood using QIAamp DNA Blood Midi kit (Qiagen) (Hollox *et al.*, 2005).

## DNA Extraction from Mononuclear Cells

Mononuclear cells were collected from the separated blood (Section 0). Genomic DNA was extracted from these cells using GenElute™ Mammalian Genome DNA Kit (Sigma) as per manufacturer's recommendations.

## Standard PCR Mix

Quantities per 10µl reaction:

10x LD Buffer	1.0µl
Primers (10µM)	0.5µl
Taq Polymerase	0.1µl
Water	7.4µl
DNA (10ng/µl)	0.5µl

### 10x LD Buffer:

- Tris-HCl pH 8.8	500mM
- Ammonium sulphate	125mM
- MgCl <sub>2</sub>	14mM
- 2-mercaptoethanol	75mM
- dNTPs	2mM (each)
- BSA	1.25mg/ml

## Phusion® Polymerase PCR Mix

Quantities per 10µl reaction as recommended by manufacturer (New England BioLabs):

5x Buffer (GC buffer)*	2.0µl
10mM dNTPs	0.2µl
Primers (10µM)	0.5µl
Phusion Polymerase*	0.1µl
Water	5.7µl

DNA

1.0µl

\*Buffer and enzyme from NEB

## Agarose Gel Electrophoresis

Depending upon the size of DNA molecules to be separated, agarose gels of varying concentrations were used (0.8 to 3.5%). According to desired gel concentration, the required amount of agarose was weighed and dissolved in a corresponding volume of 0.5x TBE buffer containing 0.5µg/ml Ethidium Bromide. Before applying samples to the gel, they were mixed with 2µl of loading buffer (0.5x TBE, 8% sucrose and 0.004% bromophenol blue) per 10µl of sample volume. Gel running buffer was the same as gel preparing buffer and gels were run using a constant voltage of 140V until the bromophenol blue had reached the end of the gel. 1µl of either the 100bp or 1kb DNA ladder (500µg/ml from NEB) mixed with 7µl of water and 2µl of loading buffer was also run in each experiment. After electrophoresis, the separated DNA molecules in the gel were visualized under UV light.

## Primer Design

Primers for PRT MLT1A0 and DefHae3 assay were designed by Prof. John Armour and Patricia Aldred (Aldred *et al.*, 2005) respectively. All other primers were manually chosen using sequence information from the UCSC genome browser's human genome assembly hg18 and expected products were obtained using the in-silico PCR function; predicted PCR products were further checked by BLAT searching against the human genome assembly to look for possible non-specific products;  $T_m$ s were calculated using primer3 (<http://frodo.wi.mit.edu/primer3>), and trace archives and dbSNP from the NCBI website (<http://www.ncbi.nlm.nih.gov>) were used to check for sequence variants in the primer sequences. For any given PCR, those primer-pairs were chosen which had the most similar  $T_m$ s.

### 3. *DEFA1A3* COPY NUMBERS

The primary aim was to apply the copy number measuring assays to European samples using calibrators of known copy numbers, which were also used to assess the accuracy of the assays. This chapter presents the results from the four copy number measuring assays (see Sections 2.1 and 2.2). The assays between them measure not only the total *DEFA1A3* copy number, but also the copy number of the full repeats (from PRT MLT1A0), copy number ratio of the *DEFA1* and *DEFA3* genes (from DefHae3 assay) and copy number ratio of the two Indel-5 alleles. It is assumed that each repeat, full and partial, carries one copy of the *DEFA1A3* locus, and that the *DEFA1A3* locus does not exist elsewhere independently. It is also assumed that each chromosome 8 carries exactly one partial repeat and a variable number of full repeats. This means that measures from PRT DEFA4-406, which amplifies across the gene itself, should be two more than the measures from PRT MLT1A0, which is specific for the full repeats. The ratios from both Indel-5 and DefHae3 assays should correspond to total *DEFA1A3* copy numbers, because the Indel-5 PCR amplifies across a sequence common to both full and partial repeats, and the DefHae3 assay amplifies a sequence from the gene. The total copy numbers were arrived at using the MLCN program (see Section 2.4). The *DEFA3* and Indel-5 allele copy numbers were assigned manually.

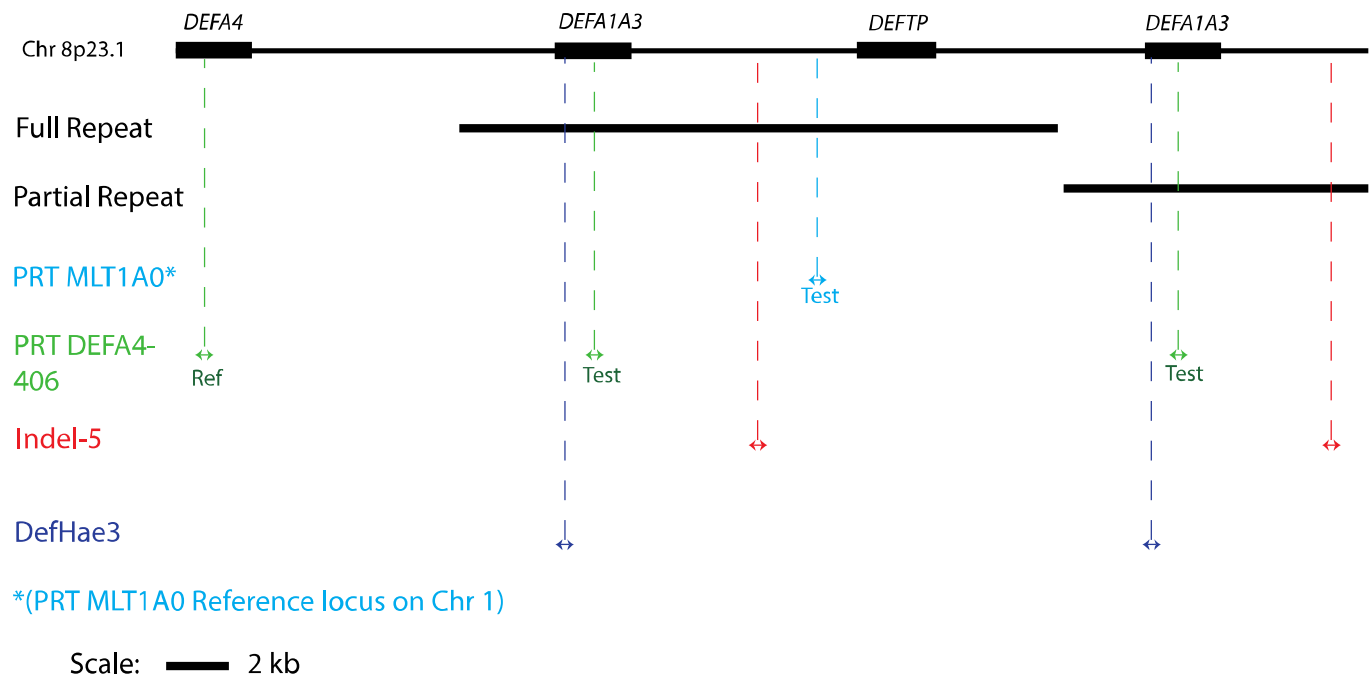


Figure 3.1 Scaled diagram of the full and partial repeats showing the position of the four assays with respect to the *DEFA1A3* locus. Note that the test sequence amplified in PRT MLT1A0 is only present in the full repeat.

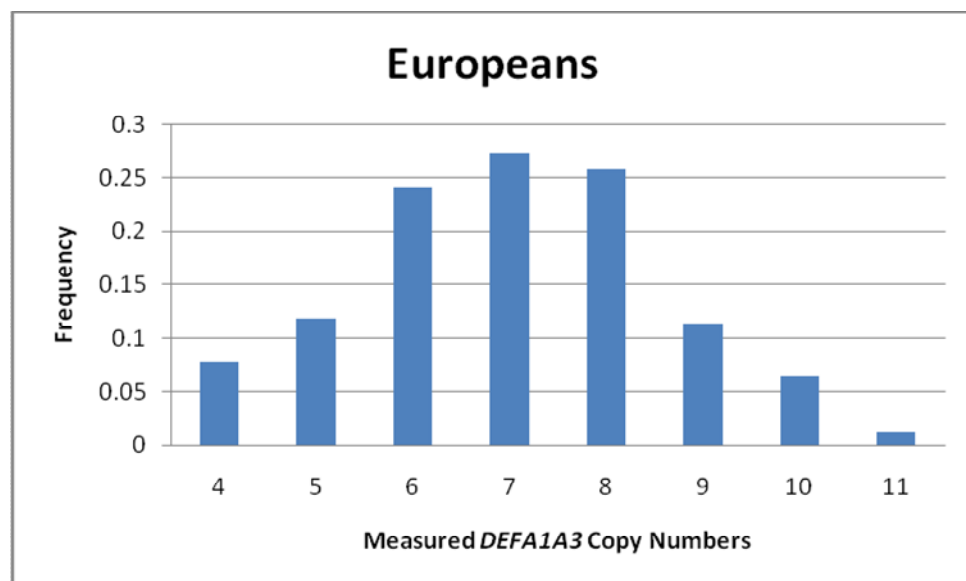
### 3.1. Total Measured Copy Numbers

Samples from three populations were typed for *DEFA1A3* copy number using the PRTs and allele ratio assays: European, African (Yorubans from Ibadan, Nigeria) and Asian (Han Chinese and Japanese). From the results of these measurements and study of haplotypes, it is clear that the substructure of this locus is different enough between these populations to treat them as three distinct groups rather than group them together. Also, the quality of the assays in terms of agreement between them for any given sample varied between the three groups. This could be due to any one or a combination of the following reasons: a) sequence divergence at this locus between the populations that affects one or more of the assays, and the use of European samples as calibrators, b) differences in DNA preparation that affect one or more of the assays. The first of these issues could be overcome by using samples from the same population as calibrators if the sequence divergence within the population is not significant or if it does not affect the PCR assays concerned. Similarly the second possible cause of error could be overcome by using calibrating samples prepared using identical procedures, and such that the procedures are consistent in the quality of DNA they produce. In this section results from measurements using the standard European calibrators are presented.



## Europeans

284 DNA samples from unrelated individuals (99 CEPH and 185 ECACC samples) were typed with the four assays. One sample carried a rare deletion of the *DEFA4* gene, which acts as the reference locus in PRT DEFA4-406, but no other variation affecting any other assay in these samples was determined. The deletion in *DEFA4* was determined by a three-primer assay (see Section 2.10) after observing the double or higher ratio of test to reference from this PRT compared to the MLT1A0 PRT in not just one sample, but also in the progeny of the sample that had also been typed. Some of the samples were tested for a deletion in the *DEFA1A3* gene itself on suspicion of getting lower than expected copy numbers from the DEFA4-406 PRT but no deletion was found (see Section 2.5). There is a 580 bp deletion catalogued as rs71776817 on the SNP database. Similar to Aldred et al's MAPH measurements, the copy numbers for *DEFA1A3* in Europeans ranged from 3 to 11. The frequency and range are depicted in Figure 3.2. Note that this figure does not include the 3-copy individual as it is the child of two samples already included. However, it signifies the presence of a one-copy haplotype in one of the samples.

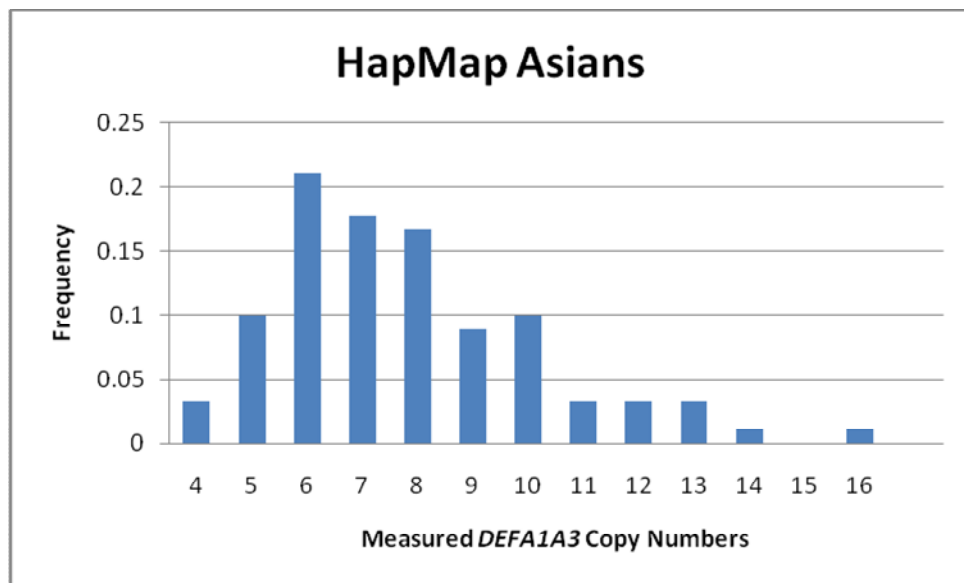


**Figure 3.2** The range and frequency of measured copy numbers in 284 unrelated Europeans.

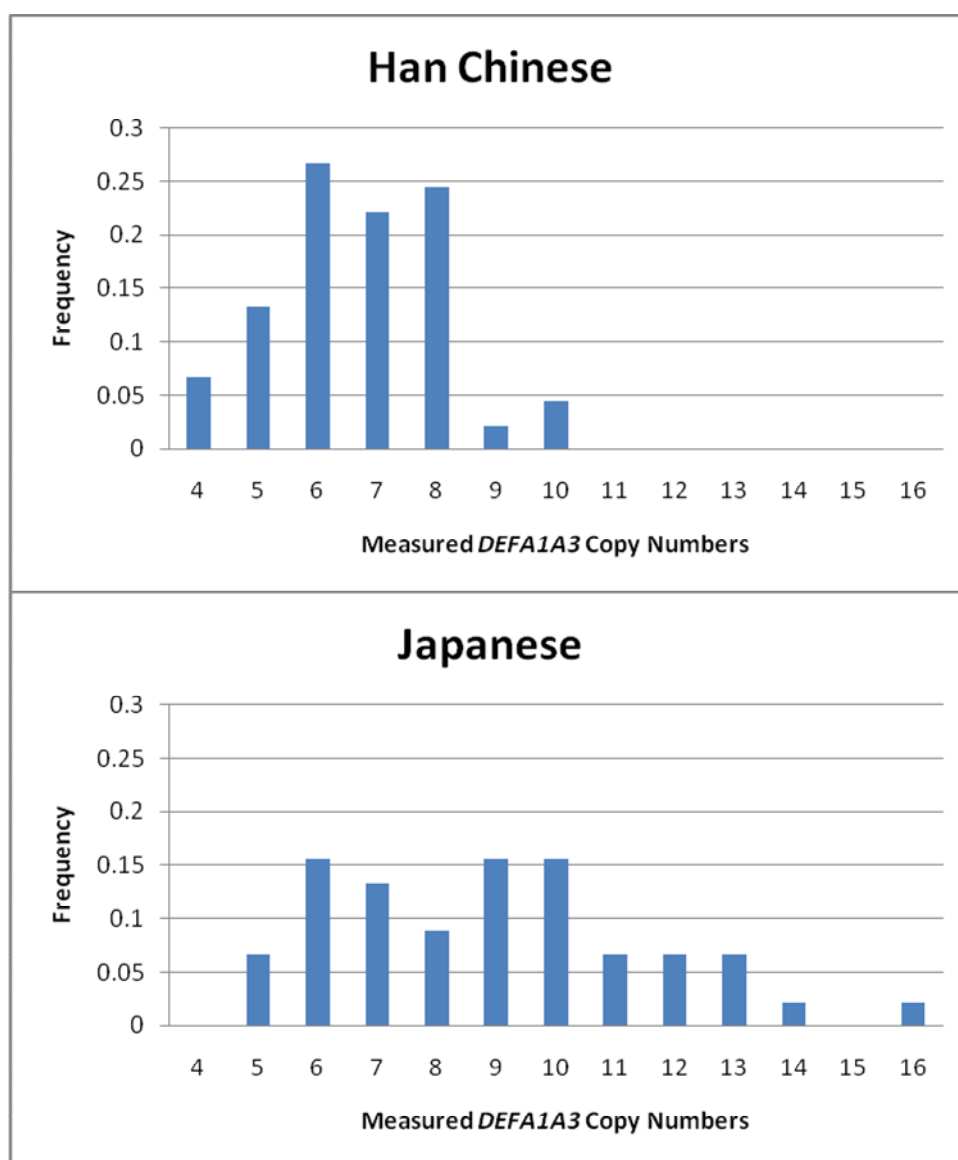
## Asians

90 unrelated DNA samples from the HapMap Asian sample set were typed with the four assays. 45 were of Japanese origin and 45 of Han Chinese. The distribution of copy numbers is shown in Figure 3.3. When the two populations

are separated, it is seen that the copy number distribution is different between them (t-test p-value=  $1.4 \times 10^{-14}$ ) (Figure 3.4).



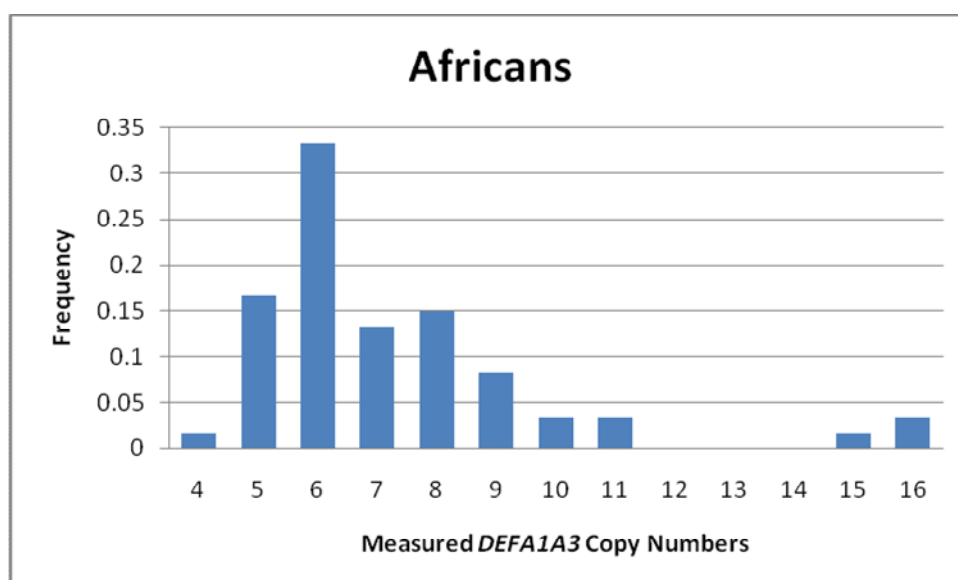
**Figure 3.3** Copy number distribution of *DEFA1A3* in 90 unrelated Asian samples



**Figure 3.4** Copy number distribution varies between the 45 Han Chinese and 45 Japanese samples typed.

## Africans

60 unrelated Africans from the HapMap Yoruban collection were typed with the four assays. The range and frequency of the measured copy numbers is given below (Figure 3.5).

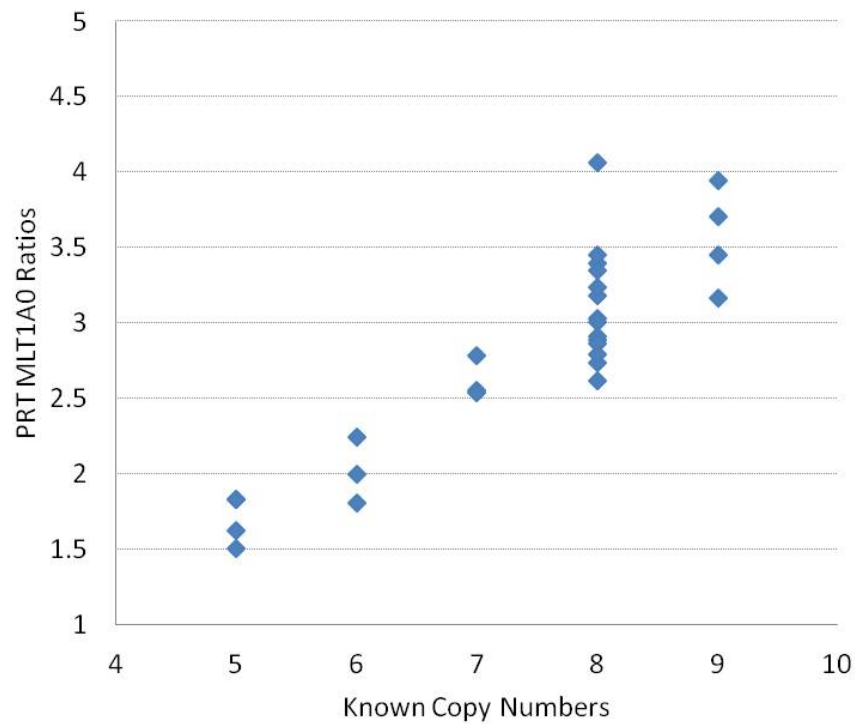


**Figure 3.5** Copy number distribution in 60 unrelated Africans.

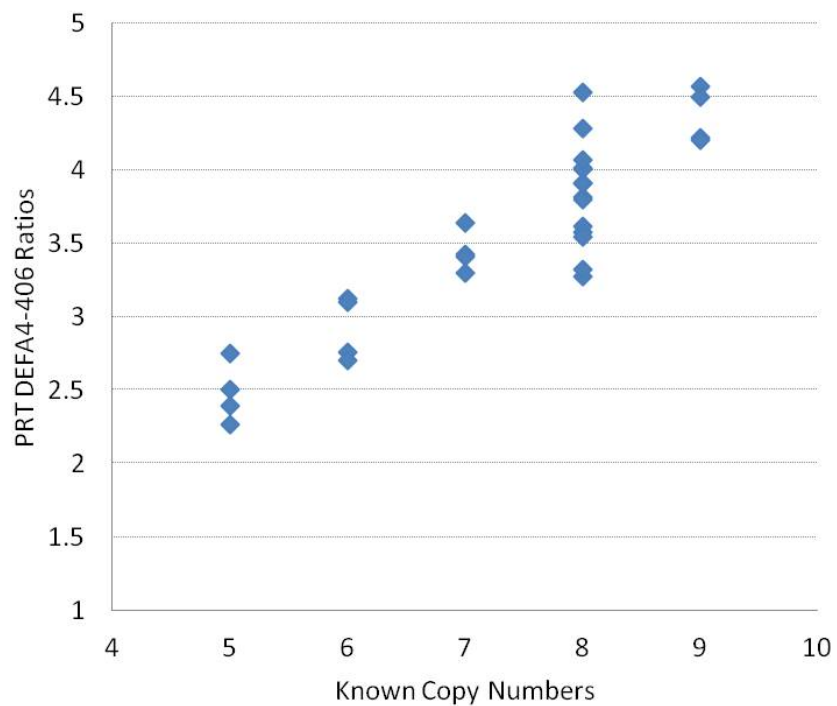
## 3.2. Copy Number Measurement Quality

### Calibrating the PRTs

As mentioned in the Methods section of this thesis, the eight samples used to calibrate the PRTs were of known copy numbers as determined by RFLP, pulsed-field gel electrophoresis and Southern blotting (Aldred *et al.*, 2005). These methods left no doubt about the accuracy of their copy number determination. Thus the calibration of the PRTs was set on firm foundations. Not only do these standard samples calibrate other tested samples, but they also act as a test of the accuracy of the PRTs. Repeated typing of these samples with the PRT assays and plotting the ratios obtained against their known copy numbers shows the accuracy and precision of these assays (Figure 3.6 and Figure 3.7). Of the eight calibrating samples, there are four with 8 copies, and one of each with 5, 6, 7 and 9 copies. Note that the ratios from PRT MLT1A0 (Figure 3.6) are lower than those from PRT DEFA4-406 (Figure 3.7). This is probably due to the fact that PRT MLT1A0 measures two copies less than the total, whereas PRT DEFA4-406 measures all copies. The actual ratio from a PRT is not important as long as the ratios from a given copy number are consistent and distinct from ratios from other copy numbers, such that plotting the ratios against the total *DEFA1A3* copy numbers should give a linear correlation. The ratios from PRT MLT1A0 are actually an average of the ratios from two tests: one with FAM-labelled primers and one with NED-labelled primers (see Section 2.1).

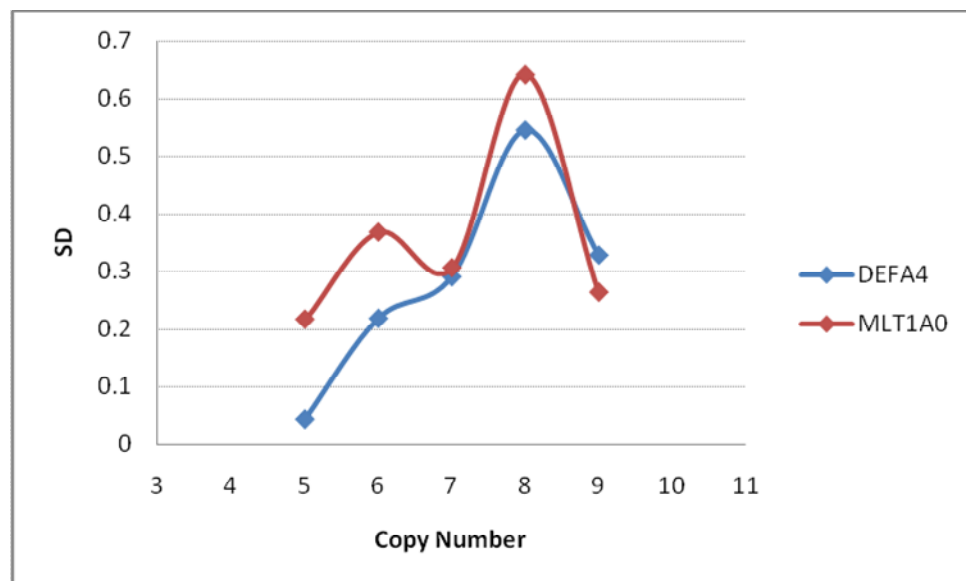


**Figure 3.6** PRT MLT1A0 ratios (averaged for both dyes) of the calibrating samples plotted against their known total copy numbers (not copy numbers of full repeats only). Each sample has been tested four times, giving 32 data points.



**Figure 3.7** PRT DEFA4-406 ratios of the calibrating samples plotted against their known copy numbers. Each samples has been tested four times, giving 32 data points.

Of the four 8-copy calibrators, there is no single one that consistently gives the higher than average ratios in both PRTs. However, it is one particular one that consistently gives lower than average ratios in both PRTs. The error in copy number calling increases as copy numbers increase, which is expected. Since PRT MLT1A0 measures only the full repeats, it is measuring 2 copies less than the other PRT for any given sample. Also, PRT MLT1A0 had the advantage of being performed twice per sample, in the form of using two separately labelled primers, as opposed to PRT DEFA4-406 which was only performed once per sample. Despite this, the standard deviation for PRT MLT1A0 is equal to or higher than PRT DEFA4-406 for the same samples (Figure 3.8).



**Figure 3.8** Standard deviations of the PRT-measured copy numbers of the calibrating samples plotted against their copy numbers.

A possible reason for this could be the nature of the sequences used in the PRTs. For PRT MLT1A0 this is a dispersed repeat sequence and thus has several copies of similar sequences around the genome. Although the primers are specific to the test and reference loci and do not match exactly anywhere else on the reference genome browser, non-specific amplification remains a possibility. Figure 3.9 shows the best-matched sequences in the genome to the test (chr 8) and reference (chr 1) loci of PRT MLT1A0 from a BLAT search. For PRT DEFA4-406 the sequence selected is an alpha defensin gene-specific sequence and only shared between the *DEFA1A3* and *DEFA4* genes.

```

Chr8      CCCAGAGAGCTCCTTCATTCATTCCCTCAGGTGAGGAAGCAGTGAGAAGGTAGCATCTAT 60
Chr6      CCCAGAGAGCGACCTTGCTCCTTCCACCATGTGAGGACACACCAAGAAGGTGTCATCTAT 60
          ***** * * * * * * * * * * * * * * * * * * * * * * * * * * * *
Chr8      GAACCAAAAAGCT-GGTCCTCACCACACACCAAATCTGTTGGCACCTTGAT-TTGGACAT 118
Chr6      GAACCAAAAAAATCAAACCTCACCAAATCCTGAAGCTGCTGGCACCTTGATCTGGACTT 120
          ***** * * * * * * * * * * * * * * * * * * * * * * * * * * * *
Chr8      CTCAGCCTCCAGAACTGAGAAATGTATTTTGTGTTTATAAGTCAC 167
Chr6      CCCAGCCACCAGCACTGTGAGAAATGTATTTCTGTTGTTTATAAGT--- 166
          * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

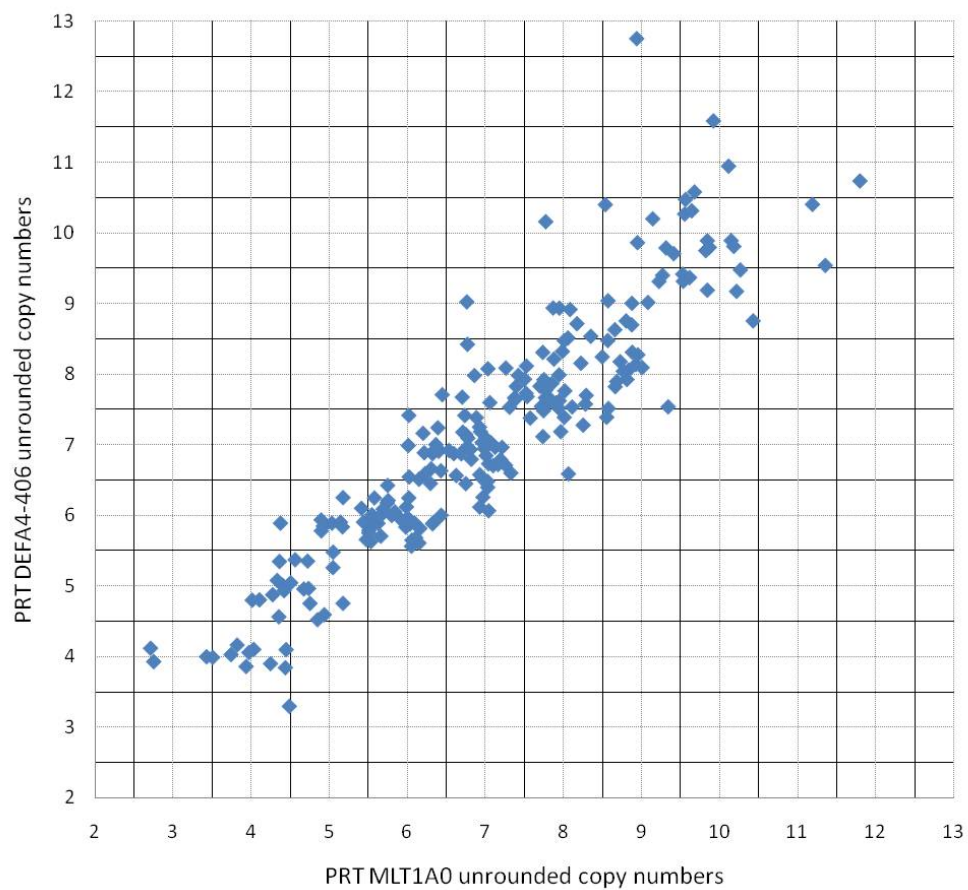
Chr1      CCCAGAGAGCTCCTTCGCACCTTCCACCATGTGAGGACACAGCAAACAGATGGCCATCTA 60
Chr1_b_   CCCAGAGAGCTCCCTAACCCCTTCTACTGTGTGAGGATGCAGCAAA--ATGGCCATCTA 57
          ***** * * * * * * * * * * * * * * * * * * * * * * * * * * * *
Chr1      TGAACCAGGAAGTGGGCCCTCACCAGATACAGGATCTGCCAGGACCGTGATCTTGGAATT 120
Chr1_b_   TGAACCAGGAAGCAGGCCCTCACCAGACACTGAATCTGTGAGAGTCTTGATCATGGACTT 117
          ***** * * * * * * * * * * * * * * * * * * * * * * * * * * * *
Chr1      CCCAGCCTCTGGAACCATGAGAAGTAAAGTTTTTGTGTTTATAAGTCAC 170
Chr1_b_   CCCAGCCTCCAGAACTATGAGAAGTAAATTTCT-GTTGTTTATAAGCCAC 166
          ***** * * * * * * * * * * * * * * * * * * * * * * * * * * * *

```

**Figure 3.9** Alignment of the best-matched sequences to the PRT MLT1A0 test locus (chr8) and reference locus (chr1). Primer sequences are in red.

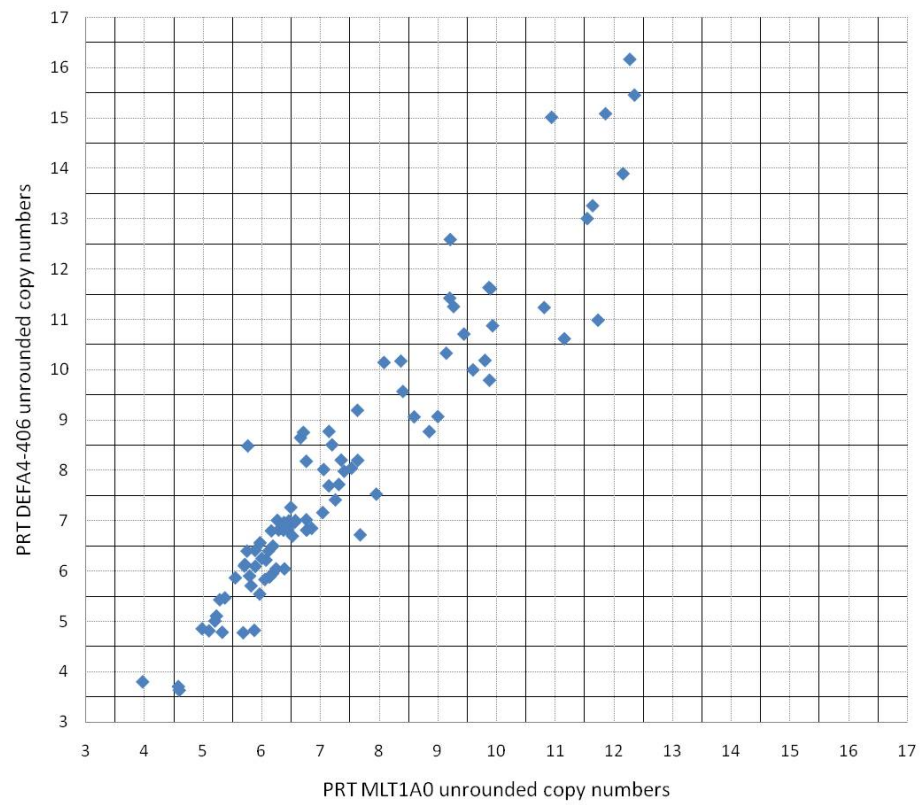
## Agreement between the PRT measurements

Agreement between the PRT measurements is evidence of accuracy for both the assays. PRT MLT1A0 measures a sequence in the full repeat outside of the *DEFA1A3* locus and PRT DEFA4-406 measures across the gene itself. However, since both PRTs are calibrated against total genomic *DEFA1A3* copy numbers, they should give the same measurement for each sample, unless there is a deviation from the assumed structure of this CNV. As mentioned earlier in this chapter, agreement between the results of both PRTs varies between the three different population samples typed. The highest agreement is seen in the European samples (Figure 3.10), in the form of data point clusters on the graph. This clustering breaks down in the case of Asian and African samples (Figure 3.11 and Figure 3.12). In the case of the discordant African samples, PRT DEFA4-406 usually gives a higher copy number measurement than PRT MLT1A0. Some almost twice as high but not quite. However, none of these samples carried a deletion of the reference sequence as tested by the *DEFA4* Deletion Assay (see Section 2.10). This decrease in agreement means that the confidence in assigning copy numbers even after taking into account the allele ratio assays, especially in the case of African samples, decreased.

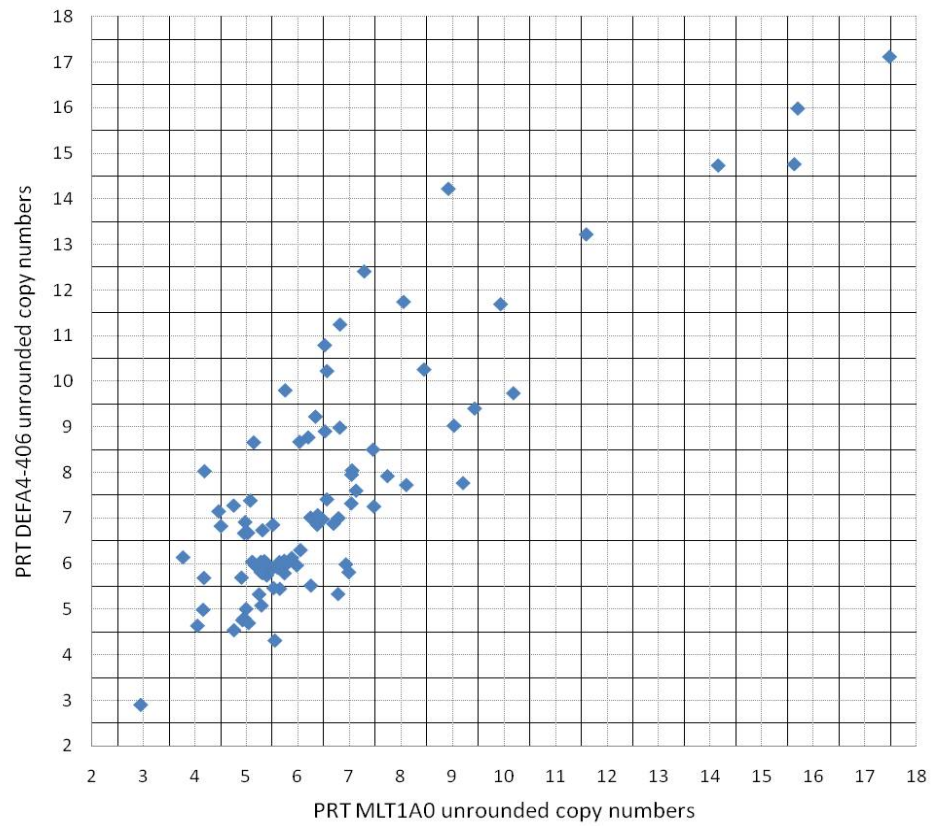


**Figure 3.10** Clusters of copy numbers observed when the unrounded copy numbers measured from both PRTs are plotted against each other. The correlation between the two is  $R^2=0.8433$  (from MS Excel). These are results from 250 European samples (HapMap CEPH and ECACC panels).





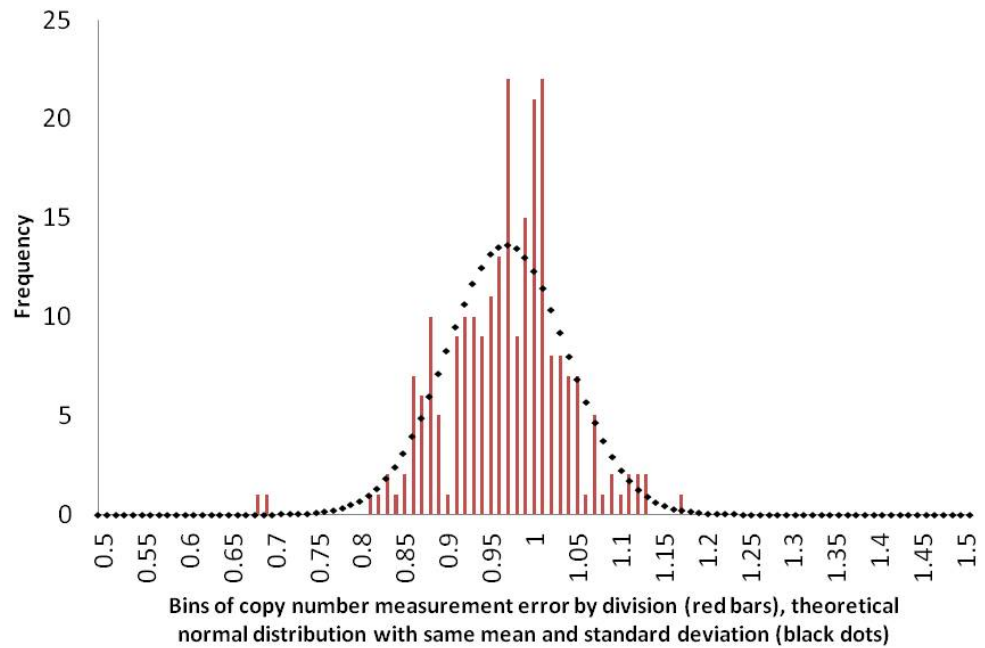
**Figure 3.11** PRT-measured copy numbers of 91 HapMap Asian samples plotted against each other (correlation  $R^2=0.8981$  from MS Excel).



**Figure 3.12** PRT-measured copy numbers of 89 HapMap African samples plotted against each other (correlation  $R^2=0.7175$  from MS Excel).

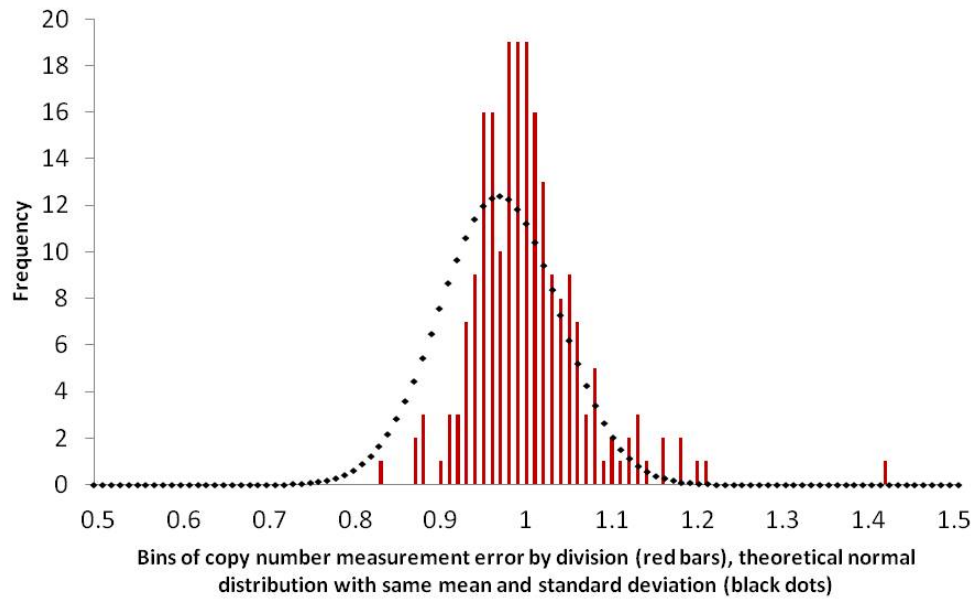
## PRT Measurement Error Distribution

Since we assume that the copy numbers of *DEFA1A3* that are measured by the PRTs are whole numbers, the measurements should centre on whole numbers and the errors should have a normal distribution if they are random rather than biased in one direction. In the histograms below are the error distributions of copy numbers from both PRTs as compared to a fitted normal distribution, using the MLCN-program generated copy numbers as references. Thus the assigned copy numbers take into account not just both the PRTs but also the ratios from Indel-5 and DefHae3 assays.



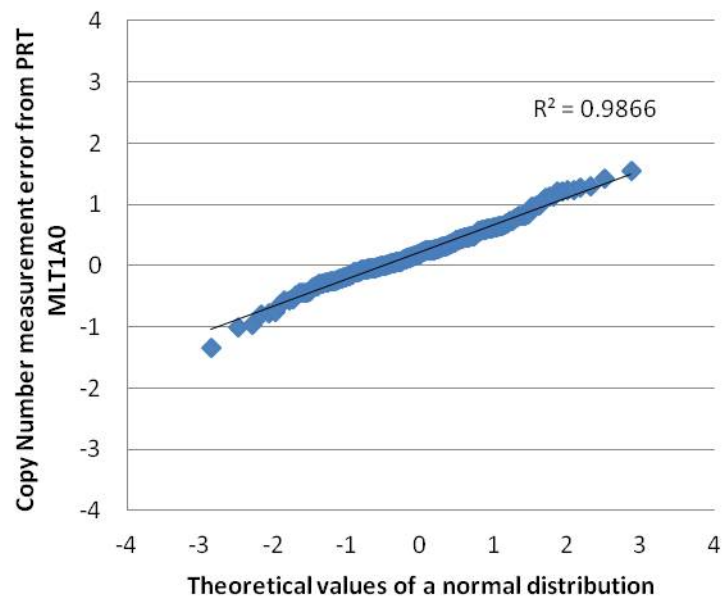
**Figure 3.13** Bars represent the frequency of PRT MLT1A0-measured copy numbers' error around a mean of 0.96 (from dividing PRT-measured copy numbers by MLCN), and the black dots represent their expected normal frequency distribution with the same mean and standard deviation.

The error for each sample is calculated by dividing its PRT-derived copy number values (unrounded) with its MLCN. These were binned into intervals of 0.01 and their histograms plotted as red bars in the figures (Figure 3.13 and Figure 3.14). Using the mean and standard deviation of these values, the number of samples to expect within each bin is calculated as the product of the normal probability distribution and total number of samples. These are plotted as the black dots in the figures.

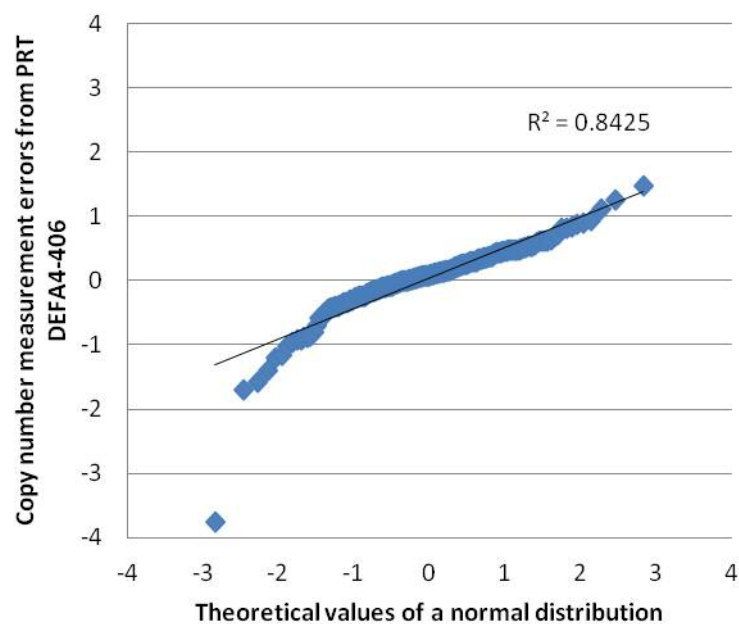


**Figure 3.14** Bars represent the frequency of PRT DEFA4-406-measured copy numbers' error around a mean of 0.99 (from dividing PRT-measured copy numbers by MLCN), and the black dots represent their expected frequency distribution.

These error distribution histograms compared with expected normal distributions are thus a visual analysis of whether there is any skew in the error of both PRTs. The QQ (Quantile Quantile) plots below are plotted to demonstrate the same, i.e. there is no significant skew in error and it follows a normal distribution for both PRTs. QQ plots are a graphical technique to check for normality, by plotting the quantiles of test sample against theoretical quantiles of a normally distributed population with the same mean. In the QQ plots below the actual error of the PRT measurement is plotted against normal theoretical values generated given the number of test samples. The error is calculated for each sample, by subtracting the PRT-measured unrounded copy numbers from the MLCN of the sample. Thus, the mean of this distribution is zero, and therefore the theoretical values were calculated in MS Excel as inverse of a standard normal probability distribution.



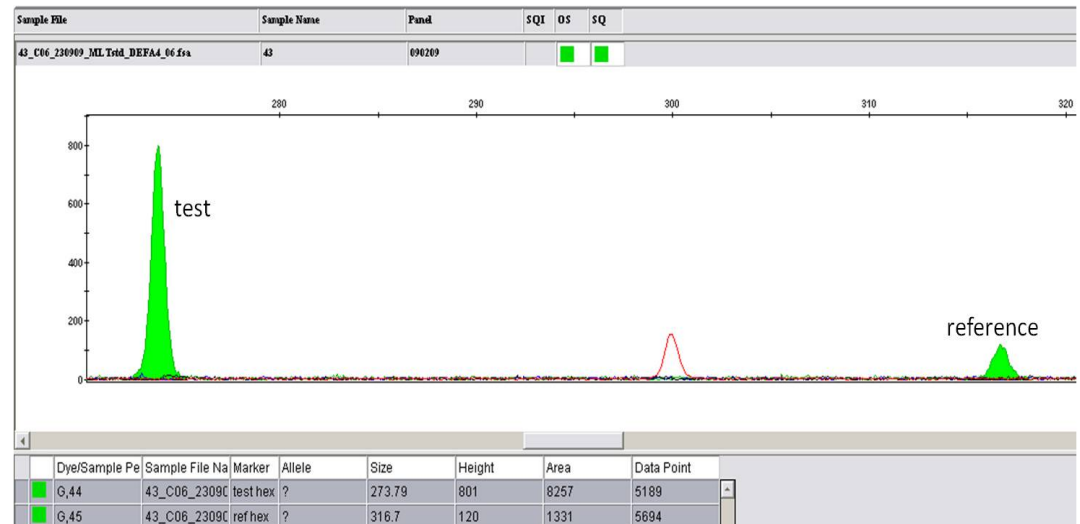
**Figure 3.15** QQ plot of PRT MLT1A0-measured copy numbers' error (from subtracting PRT-measured copy numbers from MLCN). Theoretical values are plotted on the x-axis and copy number errors on the y-axis.



**Figure 3.16** QQ plot of PRT DEFA4-406-measured copy numbers' error (from subtracting PRT-measured copy numbers from MLCN). Theoretical values are plotted on the x-axis and copy number errors on the y-axis.

The outlier in the PRT DEFA4-406 Q-Q plot is a one-off case of higher than the average error. While all other assays for this sample agree on a copy number of 9 (PRT MLT1A0 unrounded copy number= 8.9, Indel-5 ratio= 8.2, DefHae3 corrected ratio= 3.5), PRT DEFA4-406 gives an unrounded copy number of

12.7. On looking up the ABI trace for this sample, it seems that the peak from the reference product does not have the uniform shape of a good quality peak (Figure 3.17). It may be that the higher than expected ratio for this sample resulted from an error in quantifying the reference product by the ABI. The sample was not re-tested.

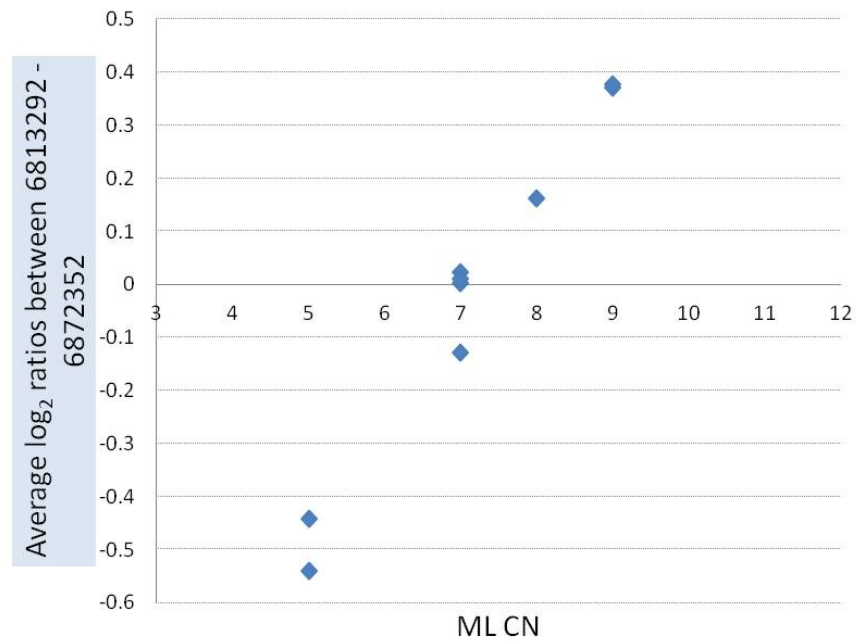


**Figure 3.17** ABI trace for PRT DEFA4-406 of a sample giving a higher than expected ratio. The peak from the reference product looks suspect.

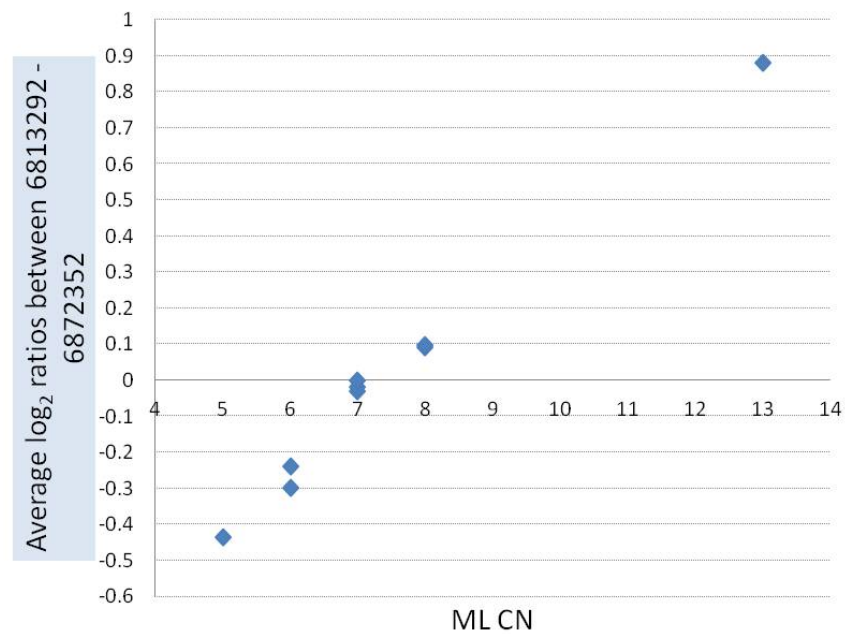
## Agreement with Microarray data

This was the only case where PRT-based copy number measurements could be compared to measurements from another method performed in another laboratory. Besides the one HapMap CEPH sample included in the samples typed by the PRTs that was used as the reference for this array- CGH experiment, 9 HapMap CEPH and 9 HapMap YRI samples also typed by the PRTs were test samples in the aCGH (Conrad *et al.*, 2010). The reference sample was measured by the PRTs to have a copy number of 7, which meant all samples that had 7 copies would give a  $\log_2$  ratio of 0 in the array CGH. Samples with copy numbers higher than 7 should give ratios larger than 0, and those with lower than 7 should give negative values.

Although it is a small number of samples but nevertheless an agreement between the two results as seen in the graphs below gives an independent source of confidence in the PRT measurements.



**Figure 3.18** ML CN plotted against log<sub>2</sub> ratios from the microarray for 9 HapMap CEPH samples. The chromosome 8 coordinates mentioned in the y-axis label are from hg18.



**Figure 3.19** ML CN plotted against log<sub>2</sub> ratios from the microarray for 9 HapMap YRI samples. The chromosome 8 coordinates mentioned in the y-axis label are from hg18.

The PRT-based ML CN and array-CGH data can also be compared via calculating the expected binary logarithms of the aCGH ratios for each copy

number. The table below shows the obtained data matches the expected values very well.

**Table 3.1** Comparison of  $\log_2$  ratios of the DEFA1A3 repeat region from array-CGH with expected  $\log_2$  ratios that have been calculated based on *DEFA1A3* copy numbers measured by the PRT-based assays

ML CN	Expected $\log_2$ ratios	Obtained $\log_2$ ratios from array-CGH (averaged in case of more than one sample)
5	-0.485	-0.47311
6	-0.222	-0.2688
7	0	-0.02028
8	0.193	0.117915
9	0.362	0.374081
13	0.893	0.88143

### 3.3. DefHae3 and Indel-5 Assay

The ratios obtained from these assays are useful for two purposes. Firstly, they corroborate the total copy numbers measured from the PRTs. Before using DefHae3 ratios, they are corrected to account for the bias caused by heteroduplex formation (see Section 2.2). To illustrate the usefulness of these ratios, consider obtaining a corrected ratio of 2.5 from a sample that has PRT-measured copy numbers between 7 and 8. This ratio would support a copy number of 7 over 8, because 2.5 can result from a ratio of 5 copies to 2 which total 7, and cannot result from a total of 8 copies. Secondly, the number of copies containing the *DEFA3* gene, the *DEFA1* gene, the undeleted and the deleted versions of the Indel-5 polymorphism are simultaneously determined. The table below lists ratios from these assays alongside the PRT-measured copy numbers for seven ECACC samples, and the interpretation of those ratios in terms of number of copies of each allele.



**Table 3.2** PRT-measured copy numbers of seven ECACC samples and the interpretation of their Indel-5 and DefHae3 ratios based on these measurements. Note that the copy numbers measured from PRT MLT1A0 have been calibrated to include the two partial repeats for each sample, and are thus copy numbers of total *DEFA1A3*, similar to the measurement from PRT DEFA4-406.

Sample ID	PRT MLT1A0 unrounded CN	PRT DEFA4-406 unrounded CN	Indel-5 ratio (deleted to undeleted)		DefHae3 ratio ( <i>DEFA1</i> to <i>DEFA3</i> )		
			measured	interpreted	measured	corrected	interpreted
10	3.50	3.99	0	4 to 0	0.91	1.05	2 to 2
11	6.98	6.94	2.51	5 to 2	4.91	5.65	6 to 1
12	7.03	8.07	3.04	6 to 2	2.41	2.77	6 to 2
14	4.73	4.96	3.89	4 to 1	1.33	1.53	3 to 2
15	5.62	5.88	1.03	3 to 3	4.31	4.96	5 to 1
16	9.61	9.36	4.03	8 to 2	3.48	4.01	8 to 2
44	8.57	9.03	3.39	7 to 2	7.36	8.47	8 to 1

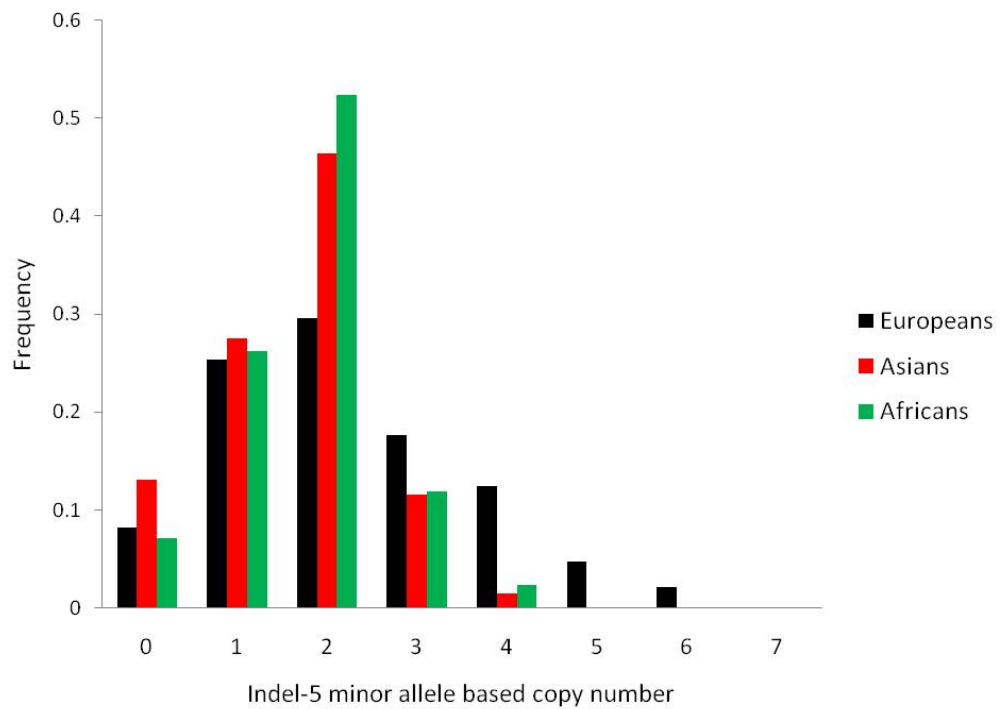
## DEFA3 Absence

It is known that *DEFA3* is absent in about 10% percent of the European population (Aldred *et al.*, 2005) and in about 10% and 37% of Asian (Chinese and Japanese) and African populations respectively (Ballana *et al.*, 2007). Using the DefHae3 ratios and the PRT-based total copy numbers, *DEFA3* gene copy numbers were determined for CEPH, ECACC, HapMap CHB/JPT and HapMap YRI samples, and similar frequencies of *DEFA3* gene absence were observed.

Population	Total unrelated samples typed	<i>DEFA3</i> absent in
Europeans	246	30 (12.1%)
Asians (CHB/JPT)	86	8 (9.3%)
Africans (YRI)	58	26 (44.8%)

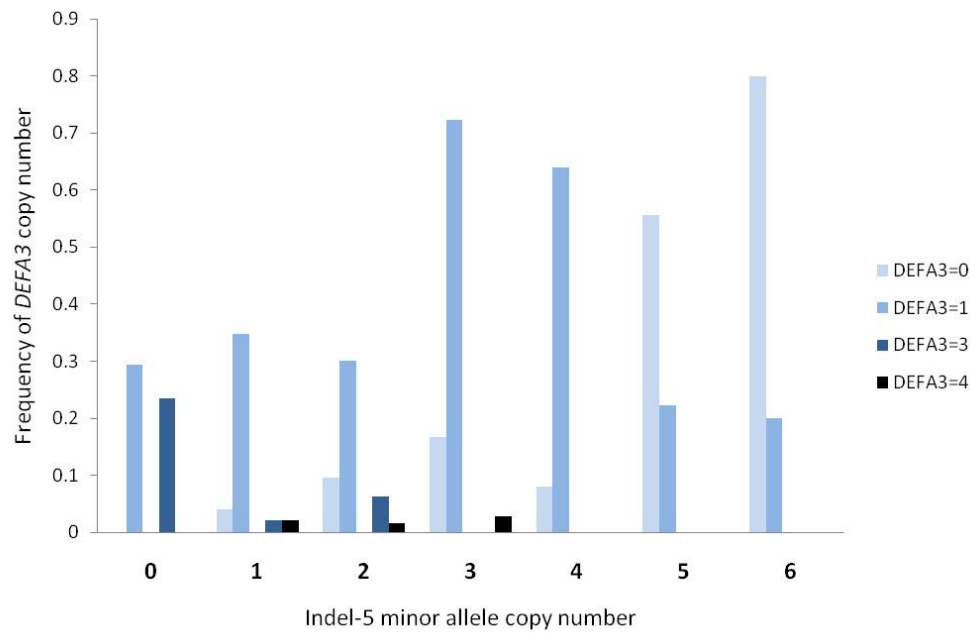
## Indel-5 minor allele copy number

As is true of total copy number, and *DEFA3* absence, the copy numbers of *DEFA1A3* repeats based on this allele also show a difference in distribution between the European and non-European populations ( $p\text{-value}=10^{-4}$ ). European samples have up to 6 copies of the minor allele whereas non-Europeans only show up to 4 copies (Figure 3.20). No difference between Asian and African samples was observed ( $p\text{-value}=0.76$ ).



**Figure 3.20** Copy number frequency of the Indel-5 minor allele in the three populations

Interestingly, a relationship between the Indel-5 minor allele and *DEFA3* gene copy number was observed in the European population. Samples with zero copies of *DEFA3* tended to have higher copy numbers of Indel-5 minor allele and vice versa (Figure 3.21).



**Figure 3.21** Frequency of *DEFA3* copy numbers for each copy number category of Indel-5 minor allele. Samples with high copies of the Indel-5 minor allele have a higher frequency of *DEFA3* absence. There is no overlap of absence of both the minor allele and *DEFA3* gene. The *DEFA3* CN=2 data have been deliberately omitted for a clearer picture.

The Asian and African samples do not show this relationship between Indel5 and *DEFA3* copy numbers.

### 3.4. Conclusion and Discussion

As has been mentioned in the introductory chapter, absence of reliable typing systems for a complex and multiallelic CNV like the alpha defensin CNV has left a gap in studying and understanding such loci in the genome. This chapter has shown how the combination of four PCR-based assays (PRT MLT1A0, PRT DEFA4-406, Indel-5, and DefHae3) can be used as a reliable and informative CNV measuring system for *DEFA1A3*. Copy numbers measured range from 3 to 11 in Europeans and 4 to 16 in Africans and Asians (Chinese and Japanese). The accuracy of the CNV measuring system has been shown by correlation between PRT ratios and known copy numbers of calibrators, agreement between the four assays, normal error distribution for the PRT-based copy numbers and correlation of measured copy numbers with an external, independent study. Since the PRT is in itself a relative measure, providing ratios instead of whole copy numbers, the importance of reliable calibrators cannot be overstated. For the *DEFA1A3* measuring system, the copy number

of the calibrators has been determined by pulsed-field gel electrophoresis (Aldred *et al.*, 2005), leaving no doubt of their correctness. Also, importantly, prior to their use for calibrating other samples, they provide a means of measuring the accuracy of the PRTs (Figure 3.6 and Figure 3.7). All four assays are PCR-based and require capillary electrophoresis for analysis of products. The PCRs on average take 2 hours and electrophoresis takes 45 minutes per 16 samples. Thus, it is a high-throughput system to perform. The designing and optimizing of the assays is the slowest stage of the process, as is true for any new study.

The lower level of agreement between the assays for the non-European, especially African, samples could be due to differences in the quality of DNA, or could result from SNPs in those populations that are not represented in the Europeans. Now, with the availability of genomic sequence data from large-scale sequencing projects, most notable being the 1000 Genomes project, even relatively low frequency sequence variants can be checked in several populations. On checking the latest SNP databases that include data from the 1000 Genomes project, two SNPs were found in the forward primer-binding sequence of PRT DEFA4-406 reference site. One, rs7816407, has a minor allele frequency of 0.009, but no information on which population it was found in. It is also reported by another sequencing project, where they found it in an African (Northern Kalahari) sample. The other, rs2741681, has a minor allele frequency of 0.025 in the complete 1000 Genomes data, but was found in the pilot scale African (YRI) sample (59 individuals) to have a minor allele frequency of 0.102. This SNP is a C to G transversion at the 9<sup>th</sup> base of the forward primer from the 3' end. If this SNP is present at a MAF of 0.102 in Africans, it should be found in 10.2% of African individuals, and if it affects primer binding, it should result in a higher than expected measure of copy number from PRT DEFA4-406, which is about what has been observed in the tested African samples (Figure 3.12). However, no sample was checked for this SNP, and it remains uncertain what causes the discordance in the African samples. At least in Europeans, this measurement system remains robust and applicable, and can also be used for delineating haplotype copy numbers by typing three-generation families as has been done in the next part of this study (Chapter 4).

## 4. *DEFA1A3* HAPLOTYPES

In this part of the study the aim was to obtain copy number haplotypes for *DEFA1A3*, obtain mutation rates (copy number changing mutations) and combine haploid copy numbers with surrounding SNP genotypes by using the copy number measuring assays to obtain diploid copy numbers in three-generation families, which had members included in the HapMap database.

### 4.1. Segregation Analysis of CEPH families

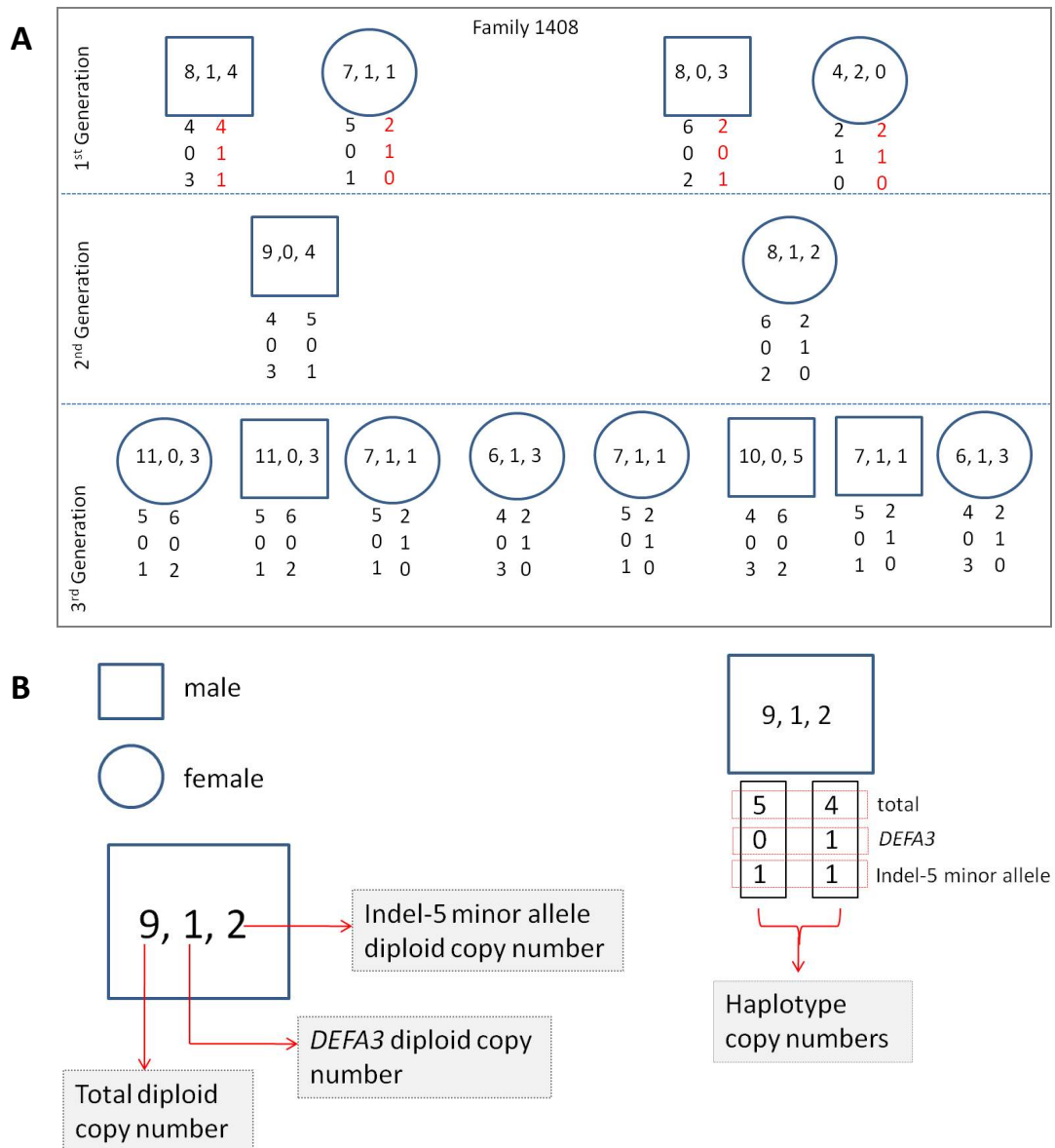
To obtain haplotype copy numbers for the *DEFA1A3* locus, DNA samples of individuals from 17 three-generation CEPH families of European origin were typed with the two PRTs (MLT1A0 and DEFA4-406) and the two allele ratio assays (Indel-5 and DefHae3). These DNA samples were obtained from the Coriell Institute for Medical Research and their details are provided in section 2.14. The details of the families are given in the table below. Of the 64 individuals of the first generation, 28 are also part of the CEU HapMap samples. They are highlighted in red.

**Table 4.1** Table detailing total number of individuals and number of individuals in each generation of the 17 CEPH families typed. Family ID is also included.

Family Number	Total Individuals	Generation		
		1 <sup>st</sup>	2 <sup>nd</sup>	3 <sup>rd</sup>
1362	17	4	2	11
1408	14	4	2	8
1341	13	4	2	7
1350	13	4	2	7
1346	14	2+2	2	8
1334	13	4	2	7
1375	12	2+1	2	7
1416	16	2+2	2	10
1345	13	2+2	2	7
1332	16	4	2	9

Family Number	Total Individuals	Generation		
		1 <sup>st</sup>	2 <sup>nd</sup>	3 <sup>rd</sup>
104	14	2	2	10
66	11	3	2	5
1333	14	4	2	8
884	18	4	2	11
1331	17	4	2	10
12	13	4	2	7
1424	14	4	2	7

These families are useful for segregation analysis because of the large number of individuals in the third generation, which allows the observation of all four possible combinations of haplotypes from the parents (2<sup>nd</sup> generation individuals) and because of the availability of segregation information from genetic markers across their genomes ([www.cephb.fr](http://www.cephb.fr)). For this study, segregation information for 44 markers around the *DEFA1A3* locus was used. Of these, 20 were on the telomeric side of the locus and 24 were on the centromeric side. The closest marker on the telomeric side is at the coordinate position of 6,741,958 which is about a distance of 73 kb and on the centromeric side is at the position 8,805,729 which is about 1.9 Mb from the locus. This segregation data tells us which two grandparents (1st generation individuals) contributed to each genetic locus typed in the third generation.



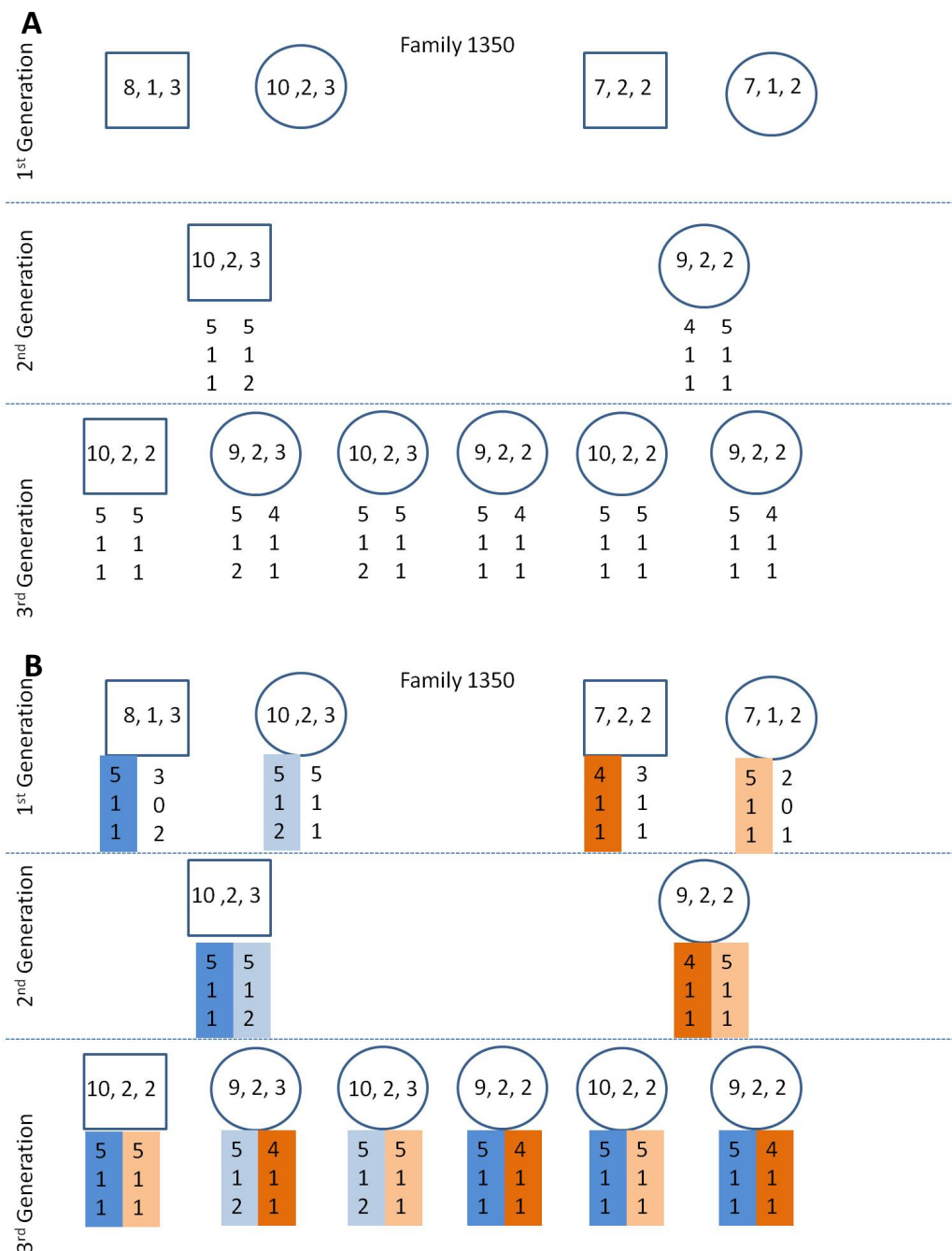
**Figure 4.1** 'A' shows family 1408, in which haplotype copy numbers for all individuals were easily inferred from their measured diploid copy numbers. This was because there was only one possible arrangement of alleles. The red coloured haplotypes in the grandparents are the untransmitted ones. 'B' describes the symbols and numbers used in 'A'.

For any given family, to obtain haplotype copy numbers of *DEFA1A3*, *DEFA1*, *DEFA3*, and Indel-5 alleles, first the diploid copy numbers from the two PRT measurements and allele ratio assays for all members were assigned manually. These were then used to infer the possible haplotype combinations in each individual. For some families no further information was needed to assign all haplotypes unambiguously. A good example of this scenario is family 1408 (Figure 4.1). The female in the second generation (mother) was inferred to have haplotypes with copy numbers 6 and 2 from her own, the father's and their children's diploid copy numbers. Given that the maternal grandmother

has a diploid copy number of 4, she cannot have passed on a 6-copy haplotype to her daughter and so it becomes clear that the 2-copy haplotype came from the maternal grandmother, and the 6-copy haplotype came from the maternal grandfather. The other two haplotypes from these grandparents were then inferred given their diploid measurements. In the case of the father in this family, the haplotype copy numbers inferred were 4 and 5. Either of the paternal grandparents could have a 4 or 5-copy haplotype as their diploid copy numbers are 8 and 7. However, the 4-copy haplotype has three copies of the indel-5 minor allele which is greater than the diploid indel-5 minor allele copy number in the paternal grandmother (only one) but fits in with the paternal grandfather's indel-5 minor allele diploid copy number of 4. Thus the 4-copy haplotype is assigned to the paternal grandfather and the 5-copy haplotype is assigned to the paternal grandmother. This segregation of haplotypes was then checked by comparing to the segregation information from linked markers in the CEPH genotype database and was found to agree with it.

For some families, it was not possible to unambiguously assign haplotypes to the grandparents on the basis of measured copy numbers alone. For example, in family 1350 the haplotypes inferred for the father are both 5 copy with one copy of *DEFA3* (Figure 4.2). This allows easy assignment of total and *DEFA3* copy number haplotypes in the paternal grandparents but the indel-5 minor allele haplotype copy numbers are different and in a manner that both haplotypes are compatible with both paternal grandparent diploid copy numbers. A similar situation for the mother and maternal grandparents is also seen in this family, where both haplotype copy numbers (4 and 5) are compatible with both grandparents and *DEFA3* and indel-5 minor allele copy numbers are also not informative. To illustrate this limitation, the grandparents' haplotypes have been left blank in part A of the figure. Using the segregation information of surrounding genetic markers from the CEPH genotype database, it was learnt which child had inherited from which two grandparents, and from this information the haplotypes could be assigned in the first generation (Figure 4.2 B).





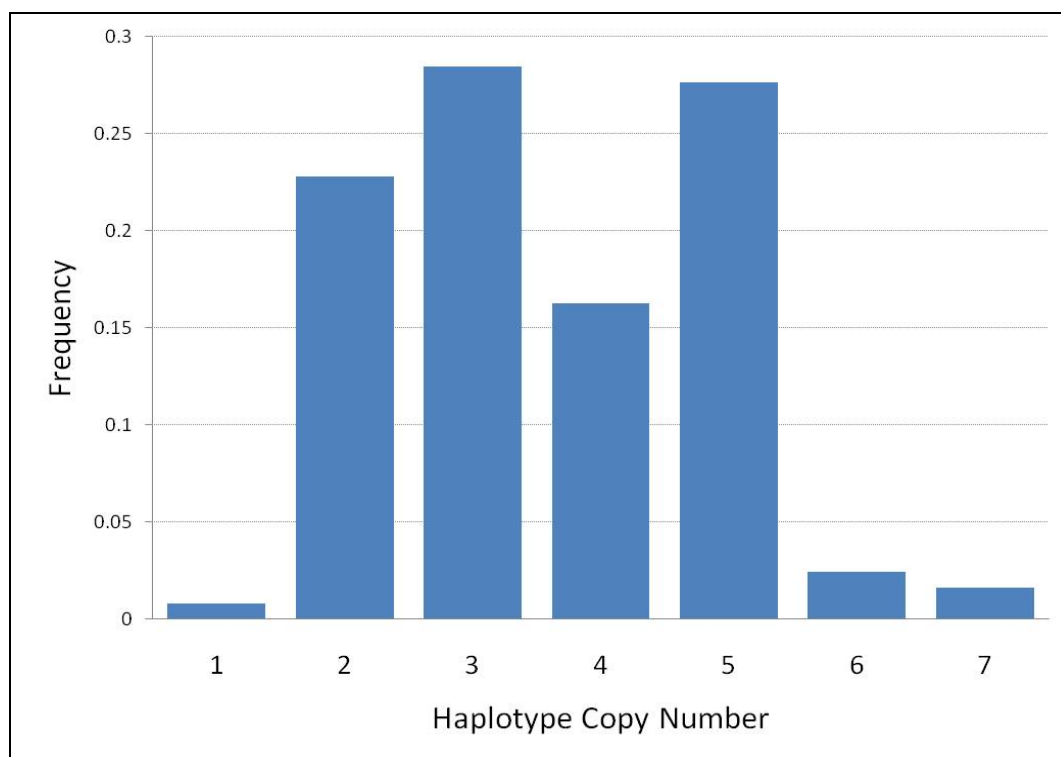
**Figure 4.2** CEPH family 1350, where haplotypes for the second generation have been inferred given the diploid copy numbers measured in generation 2 and 3. ‘A’ shows that from this information alone which haplotype comes from which grandparent cannot be inferred. All combinations are possible. ‘B’ shows the assignment of haplotypes based on segregation information of surrounding genetic markers which tells us for each child which two grandparents contributed to his/her genotype.

While the segregation information from the CEPH database is necessary to assign haplotypes in some families, they are equally necessary in others to check the completely independent haplotype assignment. That is because given the possible errors in copy number measurements, the possibility of

haplotype rearrangements during gamete formation, and not knowing the true extent of structural complexity and the rate of copy number changes at this locus, incorrect assignments of haplotype copy numbers are possible.

## **4.2. Copy Number Haplotypes in Europeans**

For each family, the number of unique haplotypes for *DEFA1A3* (or any other locus) is twice the number of individuals in the first generation. From the 17 families studied, the total number of first generation individuals typed is 64 which can give information for 128 *DEFA1A3* haplotypes. However, 4 of these 64 samples did not give analyzable PCR products from both PRTs and at least one allele ratio assay and were thus only partially informative. For these 4 individuals only one haplotype could be inferred, the one they had passed on to their child in the second generation, and the other haplotype could not be inferred because their total copy numbers were not measured. Another first generation individual, from family 1346, had results from both PRTs and allele ratio assays, but he had a *DEFA4* deletion which rendered one PRT useless; furthermore, a high total copy number from the other PRT was observed which could not be assigned with great confidence and therefore a copy number was not inferred for his non-transmitted haplotype. In all, 123 haplotype copy numbers were obtained from 64 individuals (both haplotypes for 59 individuals and only one haplotype from 5 individuals). The total haplotype copy numbers ranged from 1 to 7 in these 123 samples (Figure 4.3). Given the diploid copy number range measured in Europeans (3 to 11), this is an expected result with the exception of observing two 7-copy haplotypes.



**Figure 4.3** Total haplotype copy numbers found for *DEFA1A3* range from 1 to 7 in Europeans.

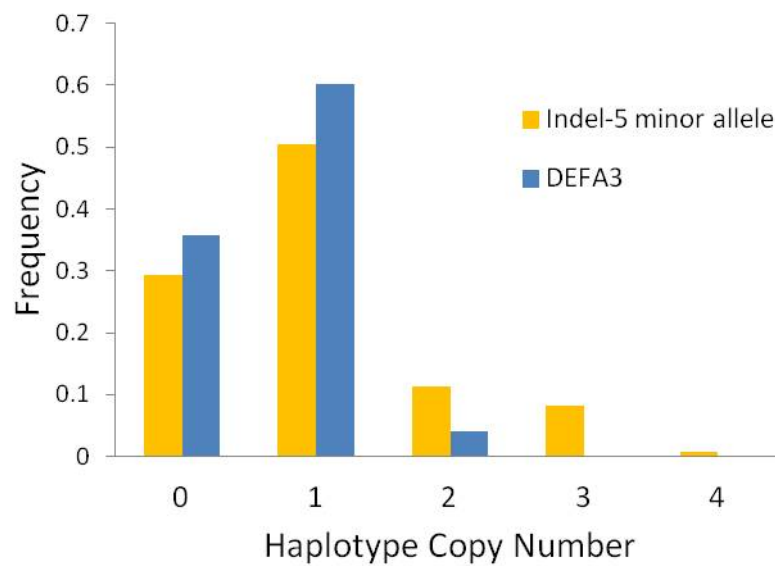
It may not be intuitively obvious why copy number haplotypes should follow this distribution (Figure 4.3) given the measured diploid copy numbers, with 4-copy haplotypes being less frequent than 3 and 5-copy ones. These are in total 123 and should be a good sample size to represent the population average. On calculating the expected frequency of each diploid copy number class from these haplotype frequencies, it can be shown that in fact this distribution matches with measured diploid copy numbers in Europeans. The table below shows these frequencies. Expected frequencies were calculated using the formula  $(a+b+c+d+e)^2$ , where 'a' to 'e' are frequencies of 2-copy to 6-copy haplotypes. 1 and 7-copy haplotypes were ignored.

**Table 4.2** Observed and expected frequencies of each diploid copy number class for European samples

Diploid Copy Number	Expected Frequency	Observed Frequency
4	0.054	0.058
5	0.136	0.093
6	0.162	0.19
7	0.229	0.229
8	0.204	0.225

Diploid Copy Number	Expected Frequency	Observed Frequency
9	0.109	0.108
10	0.088	0.081
11	0.014	0.011

The haplotype copy numbers for *DEFA3* and the indel-5 minor allele ranged from 0 to 2 and 0 to 3 respectively. These too are expected given the diploid copy numbers measured for *DEFA3* (0 to 4) and the indel-5 minor allele (0 to 6).



**Figure 4.4** The range and frequency of *DEFA3* and indel-5 minor allele haplotype copy numbers.

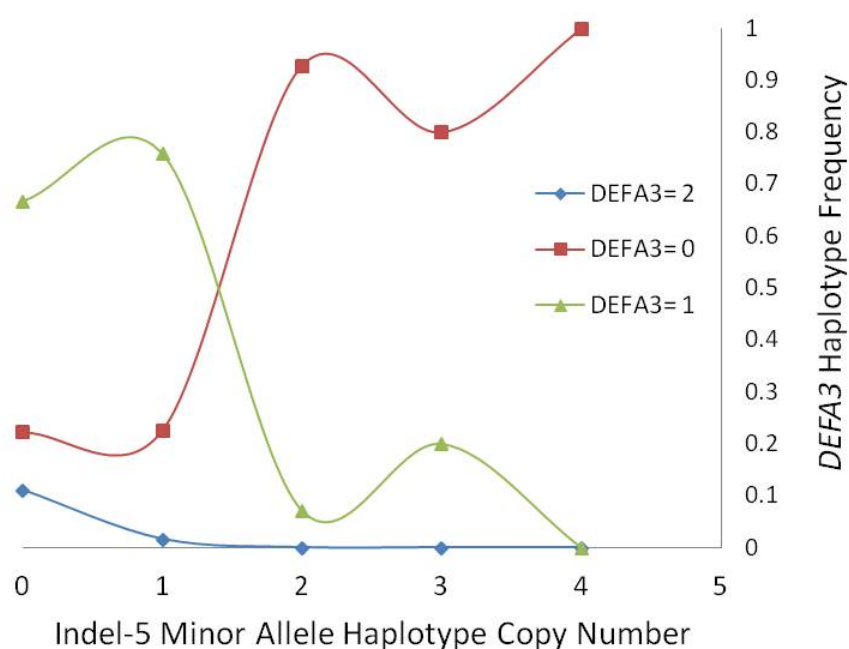
When haplotype copy numbers of *DEFA1A3*, *DEFA3* and indel-5 minor allele are combined for the 123 haplotypes, it is seen that not all possible combinations of haplotypes are observed and that some haplotypes dominate.

**Table 4.3** This table lists the copy number combinations of the 123 haplotypes inferred from the 17 CEPH families. \*\*most frequent haplotype in each copy number class, \*second most frequent haplotype in each copy number class

Total Haplotype Copy Number	DEFA3 Copy Number	Indel-5 minor allele Copy Number	Amount (% of 123 haplotypes)	% of same total copy number haplotypes
<b>1</b>	1	0	1 (0.8)	100
<b>2</b>	0	0	4 (3.2)	14.3
	0	1	5 (4)	17.8*
	0	2	1 (0.8)	3.5
	1	0	18 (14.6)	64.3**
<b>3</b>	0	0	4 (3.2)	11.4
	0	1	2 (1.6)	5.7
	0	2	10 (8.1)	28.6*
	1	0	1 (0.8)	2.8
	1	1	18 (14.6)	51.4**
<b>4</b>	0	1	1 (0.8)	5
	0	3	8 (6.5)	40**
	1	0	4 (3.2)	20
	1	1	5 (4)	25*
	2	0	2 (1.6)	10
<b>5</b>	0	1	6 (4.9)	17.6*
	0	2	1 (0.8)	2.9
	0	4	1 (0.8)	2.9
	1	1	21 (17)	61.8**
	1	2	1 (0.8)	2.9
	1	3	2 (1.6)	5.9

Total Haplotype Copy Number	<i>DEFA3</i> Copy Number	Indel-5 minor allele Copy Number	Amount (% of 123 haplotypes)	% of same total copy number haplotypes
	2	0	1 (0.8)	2.9
	2	1	1 (0.8)	2.9
<b>6</b>	0	2	1 (0.8)	33.3
	1	1	1 (0.8)	33.3
	2	0	1 (0.8)	33.3
<b>7</b>	1	1	2 (1.6)	100

As was observed for diploid copy numbers of *DEFA3* and indel-5 minor allele (see Section 3.3), an inverse relationship between the haplotype copy numbers exists for these (Figure 4.5). None of the high-copy haplotypes for indel-5 minor allele (2 and 3-copy) are also high-copy for *DEFA3* (2-copy).



**Figure 4.5** A comparison of haplotype copy numbers of *DEFA3* and the indel-5 minor allele. All 2-copy *DEFA3* haplotypes overlap with 0 or 1-copy indel-5 minor allele haplotypes.

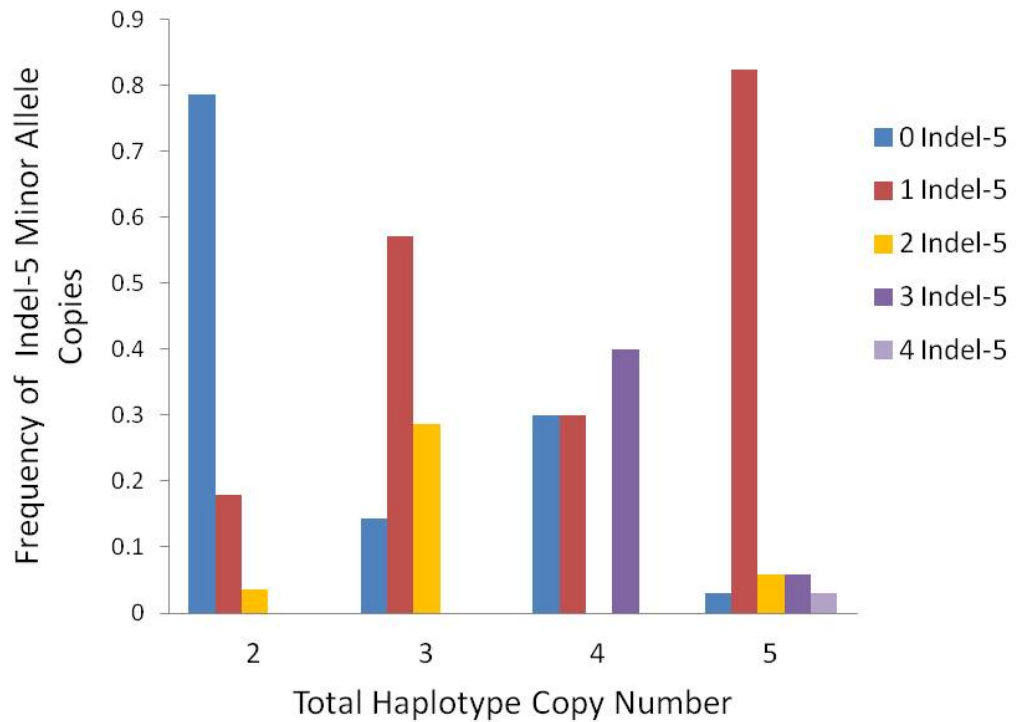
This relationship seems to be independent of total haplotype copy number, as all 2-copy *DEFA3* haplotypes have a total copy number of 4, 5 or 6, and the

highest frequency per repeat unit of the indel-5 minor allele is for haplotypes with a total of 4 copies, followed by those with 3 copies and then with 5 copies. But all 4-copy haplotypes with 3 copies of the indel-5 minor allele have either 0 or 1 copy of *DEFA3*. The same is true for 5-copy haplotypes. Also, there is an overabundance of 3-copy indel-5 minor allele haplotypes on a background of 4-copy total as shown in the table below:

Total haplotype copy number	Frequency of 3-copy indel-5 minor allele haplotypes
3	0.028
4	0.40
5	0.088
6	0
7	0

*DEFA3* is human-specific, whereas both alleles of the indel-5 variant are found in the chimpanzee genome on the UCSC genome browser ([www.genome.ucsc.edu](http://www.genome.ucsc.edu)). This fact, coupled with the haplotype copy numbers observed, supports the hypothesis that the *DEFA3* gene arose on a repeat unit which carried the indel-5 major allele. In other words, this hypothesis may explain the copy number relationship between the two. Even if this hypothesis is correct it does not mean that all existing *DEFA3*-containing repeats will have the indel-5 major allele because gene conversion events could have broken down this relationship over time. In fact, a haplotype with a total of 2 copies with one *DEFA3* and two indel-5 minor alleles has been observed in this study which makes it compulsory for *DEFA3* and the indel-5 minor allele to be present on the same repeat unit. This hypothesis can be checked by sequencing the repeat units using allele-specific primers, for example. However, it was not explored in this study.

It was also observed that the four main total haplotype copy number classes (2, 3, 4 and 5) when split according to the number of copies of Indel-5 minor allele, there are some combinations that dominate in frequency (Figure 4.6). This is especially obvious in the case of 5-copy haplotypes that largely have only one copy of the Indel-5 minor allele. The absence of 4-copy haplotypes that carry two Indel-5 minor alleles is also notable.

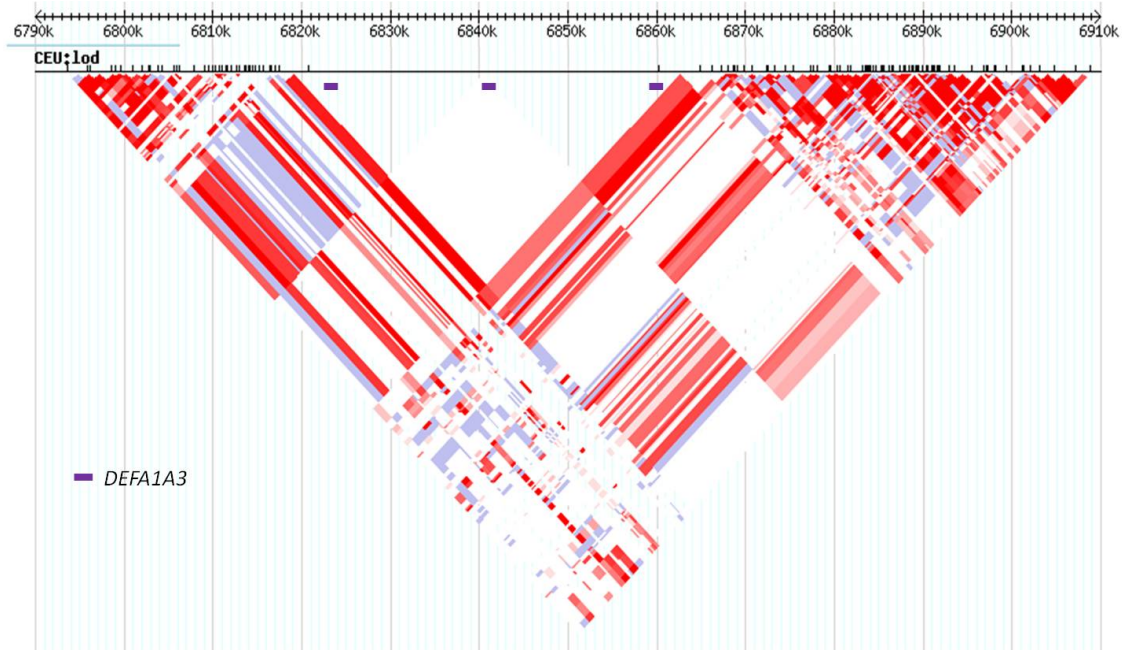


**Figure 4.6** Frequencies of each Indel-5 minor allele copy haplotype that make up each of the four main total copy number haplotypes. These are 117 haplotypes: 28 2-copy, 35 3-copy, 20 4-copy and 34 5-copy.

### 4.3. Copy Number Haplotypes and SNPs in Europeans

As mentioned in Section 4.1, of the 64 unrelated individuals from the CEPH families for whom haplotype copy numbers were obtained, 28 are also included in the HapMap database. Thus, they are also part of the HapMap CEPHs typed previously (3.1). Phased SNP genotypes for these 28 individuals across their genomes are available for analysis from the HapMap website ([www.hapmap.org](http://www.hapmap.org)). From the HapMap genome browser the LD block encompassing the *DEFA1A3* repeats spans about a 120 kb region from chromosome 8 coordinate position 6,790,000 to 6,960,000 (hg18) (Figure 4.7). This region also includes the gene *DEFA5*.





**Figure 4.7** Snapshot of the HapMap genome browser showing the LD block spanning the *DEFA1A3* repeat region, which has almost no genotyped SNPs within it.

There are 131 SNPs genotyped within this LD block on the HapMap database (data release 24, phase 2). The phased SNP genotypes for the 28 CEPH individuals were downloaded. Next, haplotype copy numbers of *DEFA1A3*, *DEFA3* and indel-5 minor allele for these individuals were added on to the SNP haplotypes. These 56 haplotypes (copy number + SNPs) were sorted in three different ways: according to total copy number, *DEFA3* copy number and indel-5 minor allele copy number. This analysis was carried out on Microsoft Excel. The copy numbers haplotypes were grouped in different ways (e.g., low versus high) and chi-squared test was used to find SNPs that differentiated between the two classes in question.

## Total Haplotype Copy Number and SNPs

The total haplotype copy numbers for these 56 haplotypes ranged from 2 to 6. Nine were 2-copy, fifteen were 3-copy, thirteen were 4-copy, eighteen were 5-copy and only one was 6-copy. The different combinations of copy number groups analyzed against each other were:

Category ID	Copy numbers in group 1	Copy numbers in group 2	Number of SNP(s) with chi-squared p-values <0.08	SNP(s) with lowest p-value (p-value)
A	2	3, 4, 5 and 6	1	rs7814783 (0.05)
B	3	2, 4, 5 and 6	5	rs4300027 and rs4512398 (0.004)
C	4	2, 3, 5 and 6	2	rs4300027 and rs4512398 (0.05)
D	5 and 6	2, 3 and 4	29	rs2978591 (0.008)
E	2 and 3	4, 5 and 6	15	rs4300027 and rs4512398 ( $1.7 \times 10^{-4}$ )
F	3 and 4	5 and 6	25	rs2978591 (0.017)
G	3 and 4	2, 5 and 6	6	rs2738048 and rs2738046 (0.042)
H	2	3	0	rs7814783 (0.185)
I	4	5	8	rs2978591 (0.019)

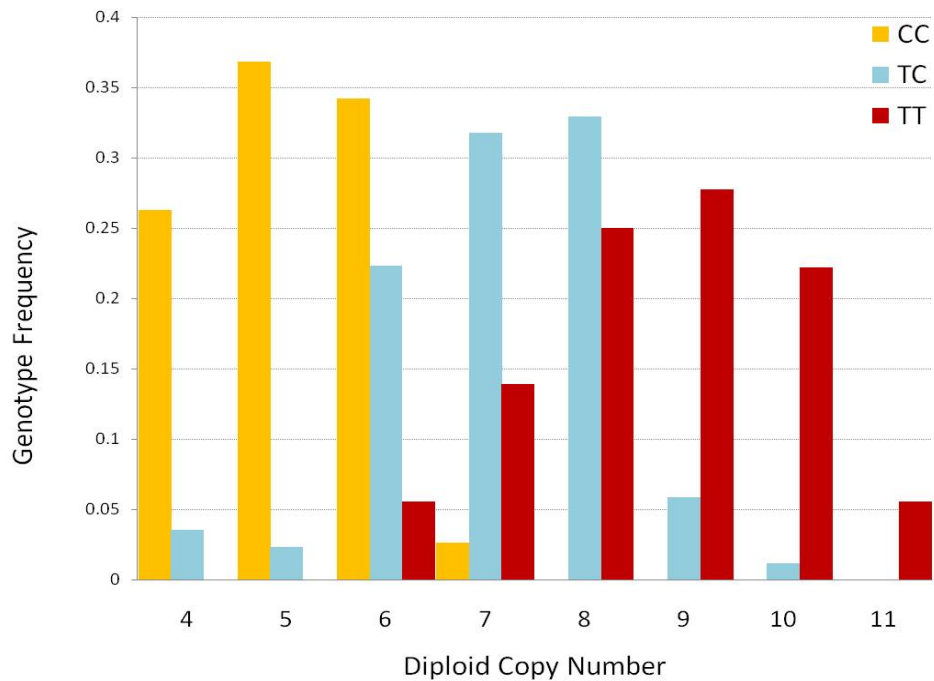
The cut-off value of 0.08 was chosen because it is low-enough, and given the average p-values obtained, it captures the lowest values in the range. Since this was a small sample size (56 haplotypes), interest was in SNPs that showed the lowest value for any given analysis, irrespective of what the value was. Further information was obtained by comparing the SNP genotypes for the high-scoring SNPs in the haplotypes in question by eye. However, using a cut-off value to count how many high-scoring SNPs are found in each analysis tells us something about the haplotypes: for example, 2 and 3-copy haplotypes differ the least from each other. Looking at the highest scoring SNPs in each category, several SNPs are shared between them. This means that some associations are only weaker replicas of others where more haplotypes have been included: for example, SNP rs4300027 is the highest scoring in categories B, C and E. This SNP has the highest chi-squared value in category E and thus the association shown in the other two categories can be interpreted

as subsets of this stronger primary association. Similarly, SNP rs2978591 is common between categories D, F and I, but shows the highest value in D.

The most interesting of all these SNPs is rs4300027 (or rs4512398, which appears to be in perfect LD with it) because of the low p-value and its ability to partition haplotypes into low and high copy number categories (2 and 3 against 4 and 5). This SNP is located just outside the centromeric end of the repeat region at chromosome 8 coordinate 6,867,985. The minor allele (C) is associated with low-copy haplotypes (2 and 3) and the major allele (T) associates with high-copy ones (4 and 5). The only 6-copy haplotype in this analysis has allele C. While none of the 4 and 5-copy haplotypes have allele C, two 2-copy and two 3-copy haplotypes have allele T. Nevertheless, this meant that if this association is real and not an artefact of small sample size, then the genotype of this SNP can be a good proxy for *DEFA1A3* copy number classes. Homozygous Cs would be expected to have diploid copy numbers ranging from 4 to 6 and homozygous Ts mainly from 8 to 10.

### rs4300027 Genotyping and Copy Number Association

To validate the finding of rs4300027 association with copy number haplotypes in Europeans, an assay was devised to genotype this SNP (Section 2.9) in the ECACC panel samples that had been assayed with the PRTs and allele ratio assays for *DEFA1A3* copy number (Section 3.1). Combining the genotypes with the measured copy numbers gave a strong association (regression analysis p-value =  $4.58 \times 10^{-28}$ ) replicating the one seen in the HapMap haplotypes (Figure 4.8). The regression was performed in MS Excel by assigning a number to denote each genotype (1, 2 and 3) and it took into account the number of samples in each copy number and genotype category.

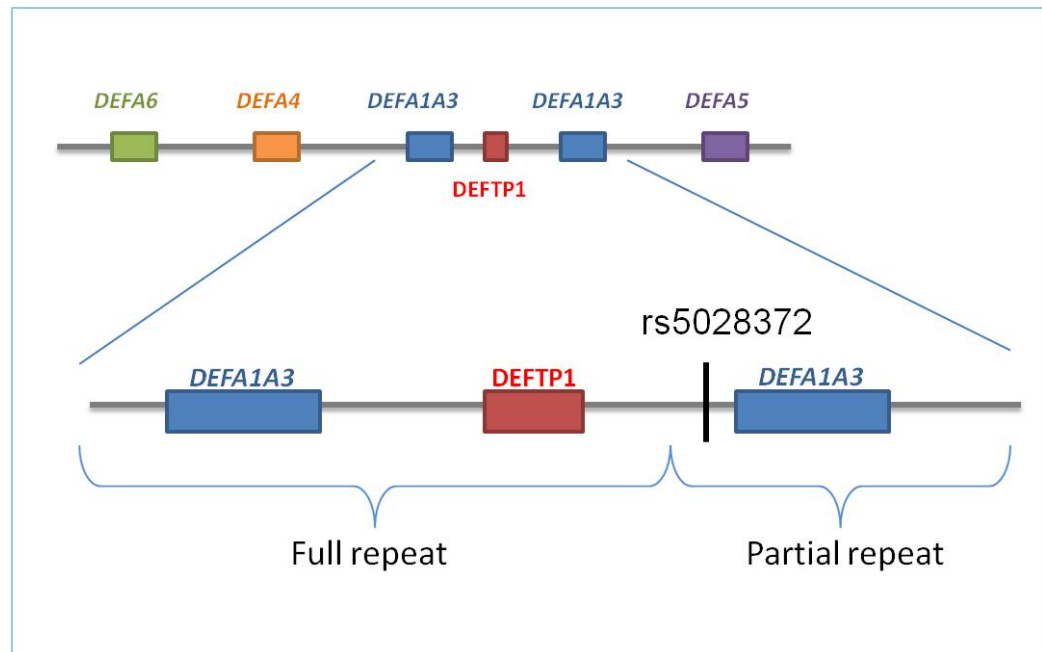


**Figure 4.8** Strong association of CC genotypes with low copy numbers of 4 to 6 and of TT genotypes with high copy numbers of 7 to 11. Number of samples in analysis is 159.

### *DEFA3* Haplotype Copy Number and SNPs

A similar analysis to that for total haplotype copy number was used to test for association between *DEFA3* copy number and SNP haplotypes for the 56 haplotypes from the CEPH families, and 24 more haplotypes from the HapMap CEPH trios (father-mother-child) were included. In these trios although haplotype total copy numbers could not be inferred, given the low *DEFA3* copy numbers and its frequent absence *DEFA3* haplotype copy numbers could be determined. Of these 80 haplotypes, 36 had none, 41 had one copy and 3 had two copies of *DEFA3*. A comparison was made between haplotypes without *DEFA3* versus those that contained at least one copy. This comparison showed several SNPs with low chi-squared p-values which suggests that presence/absence of *DEFA3* is a better preserved feature of haplotypes than total copy number, which changes more often. The highest scoring SNP in this analysis was rs5028372 with a chi-squared p-value of  $1.16 \times 10^{-4}$ . This SNP is located within the partial repeat, and is one of the six SNPs included in this analysis that are located within the repeat region. While the other five of these six SNPs are at the periphery and may or may not be in the copy number variable region, especially the three at the telomeric end

where the CNV boundary is difficult to determine, rs5028372 is well within the partial repeat, upstream of the *DEFA1A3* locus itself (Figure 4.9).



**Figure 4.9** Figure showing the position of SNP rs5028372 with respect to the *DEFA1A3* locus.

This causes concern about the validity of the SNP because the sequence in the partial repeat is almost 99% similar to that in the full repeat (on UCSC genome browser) and to differentiate between the two in practice seems difficult to achieve. Thus the question was whether the assay used to genotype this SNP was actually specific to the partial repeat or not. On the other hand, the fact that it showed a strong association in the analysis to *DEFA3* absence/presence supports its SNP status and not a multisite variant (MSV) status. On searching the HapMap database, the details of the assay used were not found.

## Indel-5 Minor Allele Haplotype Copy Number and SNPs

As was done for the *DEFA3* copy number haplotypes, four additional haplotypes from the HapMap CEPH trios were obtained for the indel-5 minor allele copy number. Thus 60 haplotypes were available for this analysis. 14 haplotypes had none, 27 had one, 8 had two and 11 had 3 copies of the indel-5 minor allele. Given the relationship observed between the haplotype copy numbers of *DEFA3* and indel-5 minor allele (see Section 4.2), it was unsurprising to find the same SNPs giving a low chi-squared p-value in this analysis as did for the *DEFA3*/SNP haplotype analysis, with the reversal of association. SNP rs5028372 is the highest scoring in this analysis as well, when

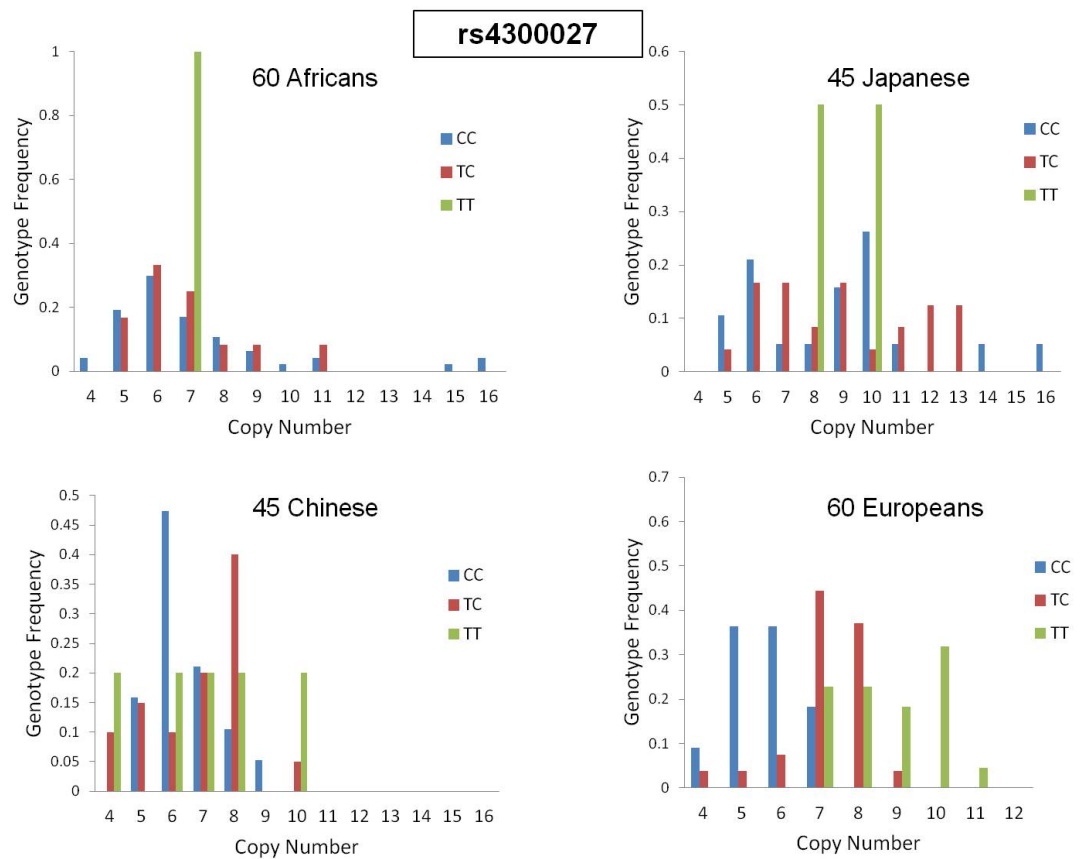
haplotypes with 0 or 1 copy of the indel-5 minor allele are compared against those with 2 or 3 copies. The minor allele of this SNP associates with 0-copy *DEFA3* haplotypes and 2- and 3-copy indel-5 minor allele haplotypes. However, for other high-scoring SNPs common between the two, highest values were obtained in this analysis when 0-copy haplotypes of indel-5 minor allele were compared against all others.

#### **4.4. Copy Numbers and SNP Genotypes in Non-European HapMaps**

Diploid measured copy numbers of HapMap African, Chinese and Japanese samples could not be resolved into haplotypes because the Asian samples were from individuals without any family members included and the Africans were only in family trios (parents and a child). They could still nevertheless be investigated for possible associations with SNPs by combining their diploid copy numbers and SNP genotypes. The genotypes for SNPs in the HapMap Asian and African samples were also downloaded from the HapMap website.

##### **rs4300027 in Non-Europeans**

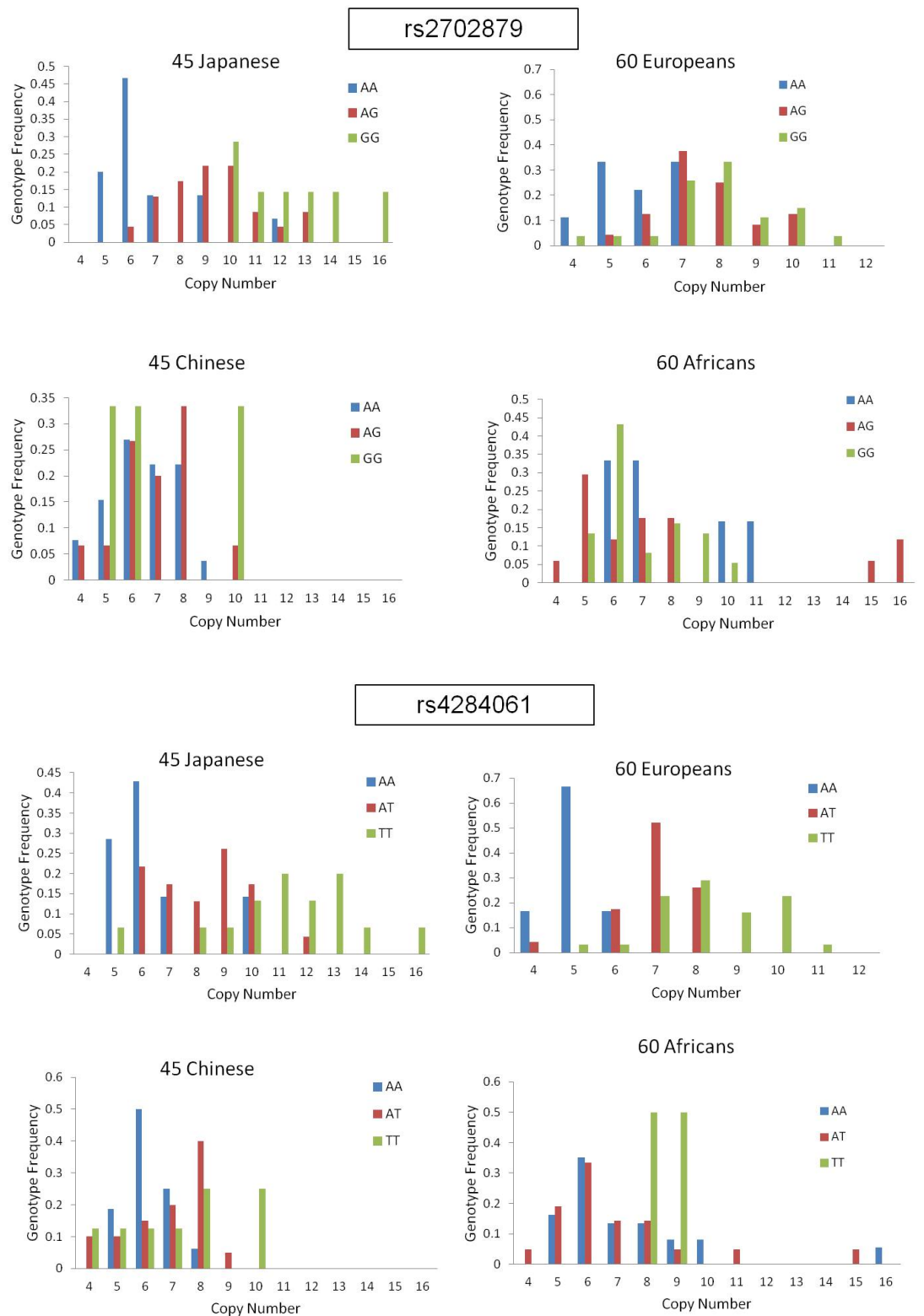
The major allele of rs4300027 in non-European populations is C instead of T (ancestral allele and major allele in Europeans is T), with the African population showing a lower minor allele (T) frequency of 0.23 (Asian MAF=0.32). The African sample did not show the association with copy number seen in the European sample and had CC genotypes with diploid copy numbers ranging from 3 to 17, and the only homozygous T individual had a copy number of 7. A similar situation was observed for the Asian samples where CC genotypes spanned diploid copy numbers from 5 to 16, and TT genotypes from 4 to 10. TC genotypes ranged from 4 to 14-copy individuals. Since the diploid copy numbers differed significantly between the Chinese and the Japanese, their copy number relationship with rs4300027 was also investigated separately. The SNP frequency doesn't change between them; the Japanese have examples of high copy numbers associated with the C allele (Figure 4.10), a pattern opposite to that observed in Europeans.



**Figure 4.10** A comparison of the distribution of rs4300027 genotypes with respect to measured diploid copy numbers in the HapMap populations tested.

## Other Copy Number Tags

While the combination of SNP genotypes with diploid copy numbers did not return any SNPs that could split the samples into copy number classes in the Africans and Chinese, there are some SNPs that show a weak tagging ability in the Japanese population. There seem to be two independent SNP blocks on either side of the repeat region that are in LD with copy number class. The two prominent SNPs are rs2702879 and rs4284061 at chromosome 8 coordinates 6,793,624 and 6,865,667 on hg18. Their associations with copy number are shown in Figure 4.11 for the Japanese and other HapMap population samples. The allele frequencies differ between the four groups: for rs2702879, allele A is at a frequency of 0.58 in the Japanese, 0.76 in the Chinese, 0.33 in the Europeans and 0.19 in Africans, and for rs4284061, allele T is at a frequency of 0.59 in the Japanese, 0.41 in the Chinese, 0.75 in Europeans and 0.2 in Africans. None of these SNPs have been studied further.



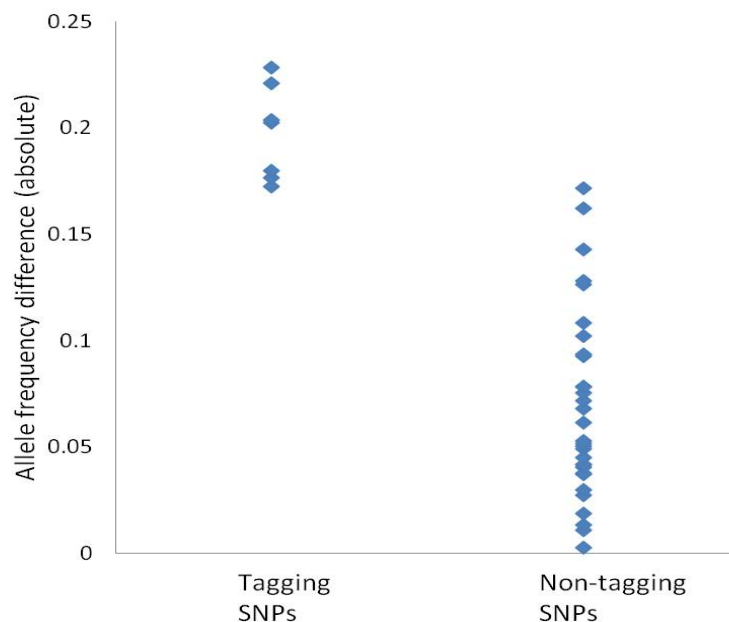
**Figure 4.11** Genotype frequencies for rs2702879 and rs4284061 in the four HapMap populations

Both these SNPs also seem to separate the Europeans into low, medium and high copy number classes. SNP rs4284061 does this because it is in weak LD with rs4300027 ( $R^2 = 0.59$ ) and shows only three possible haplotypes in



association with it. The association seen with rs2702879 is probably a subset of the association this SNP has with Indel-5 minor allele haplotypes. In the analysis done previously (Section 4.3) it is the best SNP to differentiate between haplotypes that have no copy of the Indel-5 minor allele and haplotypes that have at least one copy (chi squared p-value of 0.0069 from 60 haplotypes). Since it has been observed that Indel-5 minor allele copies have a certain association with total haplotype copy numbers (Figure 4.6) rather than being randomly associated in the Europeans, a SNP that tags Indel-5 minor allele presence/absence will likely show a weaker association with total haplotype copy number. This seems to be independent of rs4300027 ( $R^2 = 0.14$ ).

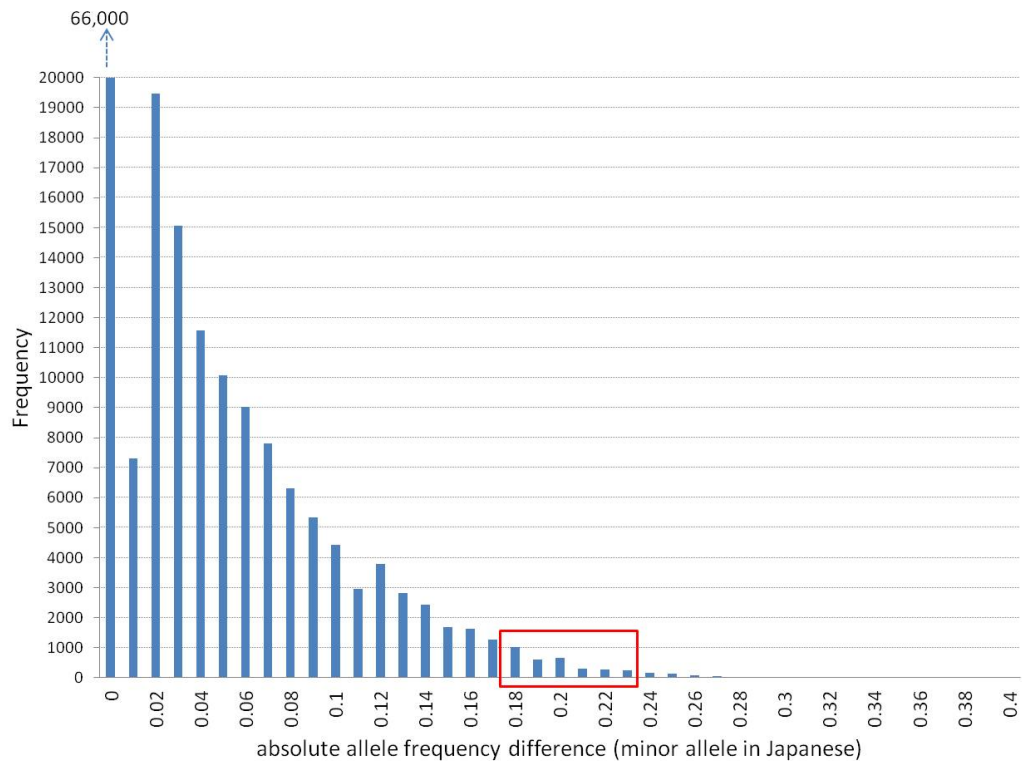
It is interesting to note that the SNPs in the *DEFA1A3* LD region that show an apparent CN-haplotype tagging ability in the Japanese have a high allele frequency difference between the Japanese and Chinese (range from 0.172 to 0.228). The other SNPs in the region that do not show this ability have a lower spread of allele frequency difference (0.002 to 0.171) (Figure 4.12). Thus the allele that has become more common in the Japanese in each case is the one associated with high copy numbers.



**Figure 4.12** Graph showing the spread of allele frequency differences between the Chinese and Japanese samples from SNPs that show a CN-tagging ability in the Japanese and those that do not

When compared to the allele frequency differences between the Chinese and Japanese HapMap samples of SNPs from the entire chromosome 8 (total

182,732 SNPs tested), the CN-tagging SNPs in the DEFA1A3 LD block lie in the top 2% (Figure 4.13).



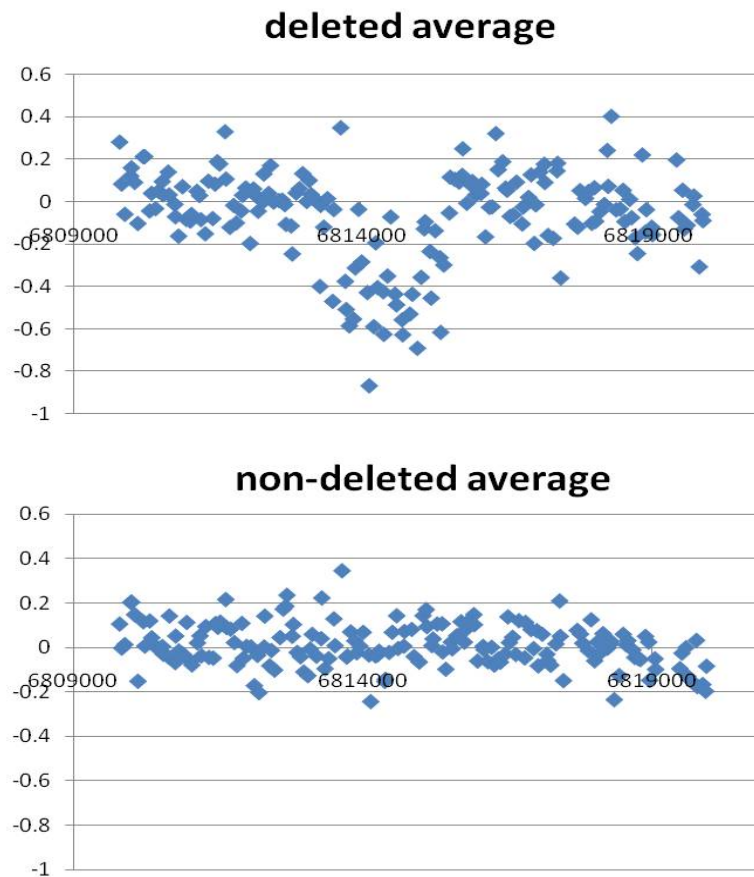
**Figure 4.13** Graph showing the number of SNPs in each allele frequency difference class, when allele frequency difference is calculated between the Chinese and Japanese HapMap samples taking that allele as minor which is minor in the Japanese. The red box highlights the classes where the CN-tagging SNPs from the *DEFA1A3* region lie.

These observations suggest that certain high copy number haplotypes have become frequent in the Japanese. The difference in the presence/absence of these haplotypes between the Chinese and Japanese populations seems quite remarkable given the otherwise genetic homogeneity. Whether this is a chance happening or a case of selection on the *DEFA1A3* locus, it needs to be investigated further.

## 4.5. Upstream Replacement Polymorphism

### Discovery and Determination

From a whole genome microarray study for CNVs (Conrad *et al.*, 2010) it was observed that four individuals had an apparent deletion upstream of the start of the *DEFA1A3* repeat region represented by a decrease in the intensity ratios (Figure 4.14).

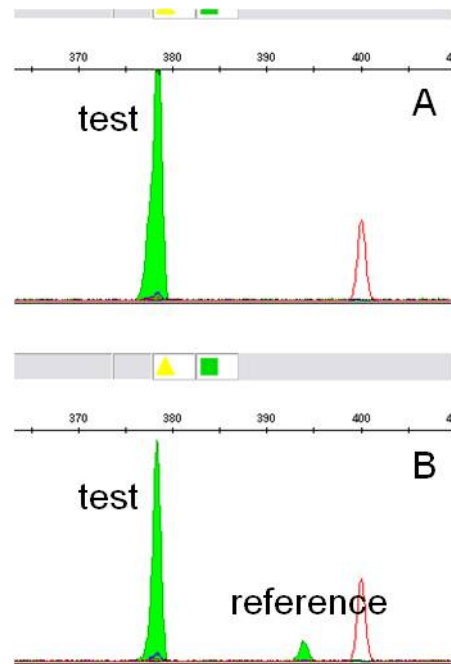


**Figure 4.14** The average  $\log_2$  ratios of the four samples exhibiting decrease in ratios and the average of 16 samples not exhibiting it. X-axis represents the probes covering the chromosome 8 region from position 6,809,795 to position 6,819,949 (hg18). Y-axis is the average  $\log_2$  ratios from the microarray experiment.

Since it is in the region in linkage disequilibrium with the repeat region, it was of interest to be able to type this polymorphism and see if it tagged any copy number lineages (Figure 4.15). From Perry et al's study (Perry, Ben-Dor *et al.*, 2008) 2 additional HapMap CEPH samples, 3 HapMap Chinese samples and one HapMap Japanese sample also presented with this deletion. Whereas the CEPH and Chinese samples had similar decreases in signal intensity ratio (binary logarithm about -0.9), the Japanese sample had a binary logarithm of -5 and seemed to be homozygous for the deletion. Primers were designed to get a short product from samples that carried the deletion. However, all samples continued to give a full-sized product. Another observation for this sequence was its similarity to the *DEFTP1* sequence in the full repeat. This sequence had a 16 bp deletion with reference to the *DEFTP1* sequence and primers were designed to amplify across this distinguishing 16 bp that would bind to both this sequence and the *DEFTP1* sequence, thus giving a third PRT for the *DEFA1A3* region (PRT *DEFTP1*, see Section 2.11).



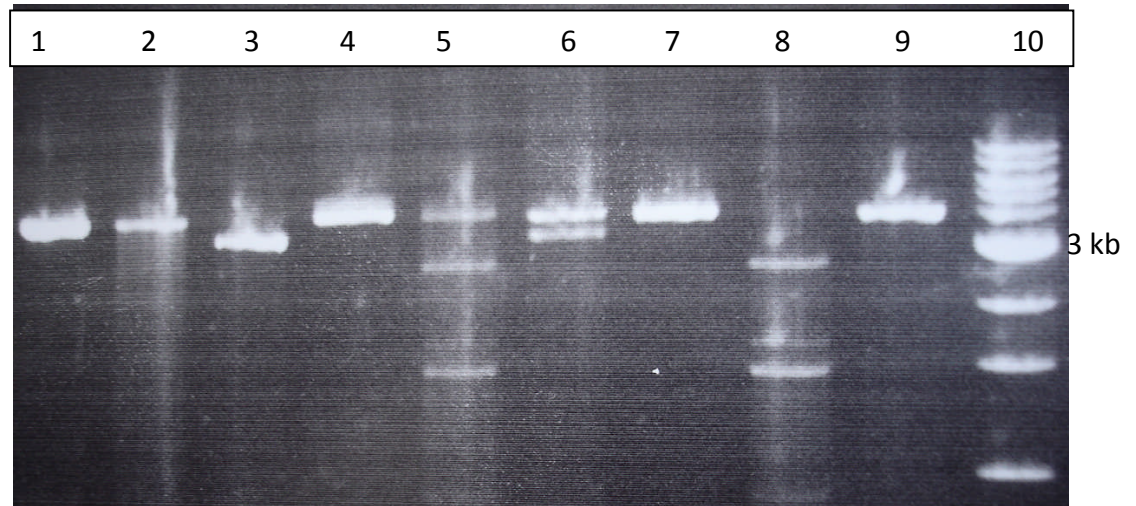
For reasons not clearly understood, there was a strong bias in the amplification efficiency of the test and reference loci and the *DEFTP1* PRT was not pursued further. However, it was useful in showing the presence of a single copy of reference sequence in the samples showing a deletion on the microarray, by giving test to reference ratios and hence copy numbers calculated from them double than measured from previous assays. It also showed absence of the reference sequence in the sample with the apparent homozygous deletion (HapMap Japanese from Perry's data) (Figure 4.16 A).



**Figure 4.16** 'A' shows the electrophoretogram of the PRT products from the HapMap Japanese sample with a homozygous deletion on the array. Only a test product-sized peak is obtained. 'B' shows the products from an unchanged sample. It has both test and reference sized products.

The fact that there was no short product with the primers designed to assay this deletion, it was hypothesized that it is not in fact a deletion but a replacement of the sequence with the paralogous sequence in the full repeat (a gene conversion event). This also explained the absence or reduction of the reference peaks of *DEFTP* PRT in samples with a decrease in microarray ratios. This was tested by identifying sequence differences between the upstream and full repeat sequences that gave rise to restriction enzyme recognition sites in either of the two but not both. Then a primer was designed outside the region of sequence similarity, further upstream, and one within. These primers are called usDEL5-F and usDEL1-R. The products were digested with two enzymes. One enzyme (*Bst*UI) would only digest if the sequence came from the original upstream sequence and the other (*Eco*RI) would digest if the

sequence came from the full repeat (Figure 4.17 ). This and another test where the upstream sequence was specifically amplified and the products used to seed a second PCR reaction that used the PRT primers, thus distinguishing the presence of the original or full repeat sequence at the upstream site by a size difference of 16 bp (data not shown), conclusively showed that it is a replacement polymorphism.



**Figure 4.17** Agarose gel electrophoresis of undigested and digested PCR products of usDEL5-F+usDEL1-R PCR. Lanes 1, 2 and 3 are from an ECACC sample which gives a ratio equal to the expected ratio from PRT DEFTP and should only be digested with *Bst*UI. In lane 1 is the undigested product, in lane 2 is product from *Eco*RI digestion and in lane 3 is product from *Bst*UI digestion. Lanes 4, 5 and 6 are from an ECACC sample which gives a ratio double the expected ratio from PRT DEFTP and should be partially digested by both enzymes. Lane 4 is undigested, lane 5 is digested with *Eco*RI and lane 6 is digested with *Bst*UI. Lanes 7, 8 and 9 are from the HapMap Japanese showing a homozygous deletion on the microarray and should be only completely digested by *Eco*RI. Lane 7 is undigested, lane 8 is digested with *Eco*RI and lane 9 is digested with *Bst*UI.

Estimating the junctions of the replacement sequence from the microarray data, a hypothetical sequence was created that partially matched the upstream sequence and partially the full repeat sequence. This was then used to BLAT search in the trace archives (NCBI database) and two exact matches were found. This was done for both ends of the replacement sequence, thus allowing definition of the junction point down to a few base pairs. The sequence around the telomeric end is given in Figure 4.18.

Chr8:6,813,079: *agcagcagatccggtataatctaccaggaagggcacaggacccaaagc  
gacgttgaaagaaatggcaaattcctcgtctgcaaatgcacctcaagcctctccctgagcctggggacac  
agggacagcatcagaaatggatcaccaaggtaacagtgggtgtaaaggaatctggagaagtcac  
gtgccagctGatgagtgatgttgGctgcattggggccagcagcatgaacagcctcagtcaataggaat  
aaatacacagagcagtgctggtcacacaggattgagactcatttcattgagctcattttgtgcttctgcc  
ccgtcacacacacagctgaacacactctgggcttggtctacttttaaaaactattctacagatacagta*

**Figure 4.18** The reconstructed sequence that had two matches in the trace archives. The blue sequence exactly matches the upstream sequence, the red sequence exactly matches the full repeat and the intervening sequence exactly matches both. The capitalized bases are differentiating ones between the upstream and full repeat sequences, such that the blue one is the last one that matches the upstream sequence and the red one is the first one that matches the full repeat sequence. Hence somewhere between these two is the beginning of the ‘replacement’ or gene conversion event.

## Haplotypes Containing the Upstream Replacement Polymorphism

Using the 3-primer assay (usDEL assay, Section 2.11), the 30 HapMap CEPH trios were typed for the replacement polymorphism. From this typing, 14 haplotypes carrying this polymorphism and 102 not carrying it were determined, while four haplotypes remained inconclusive because in one trio all three individuals were heterozygous. The status of this polymorphism was combined with copy number (where available) and SNP haplotypes. Of the 14 haplotypes carrying this replacement allele, complete copy number data was available for 5 and indel-5 minor allele copy number was available for a further 3:

Total copy number	DEFA3 copy number	Indel-5 minor allele copy number
2	1	0
3	1	0
4	0	0
4	0	0
4	0	0
?	?	0
?	?	0
?	?	1



The strongest association seems to be with the absence of indel-5 minor allele. Since this polymorphism has been observed in Europeans and Asians, but not in Africans so far, it is possible that it is a non-African variant and so quite recent. One SNP, rs4543566, about 400 bp upstream of this polymorphism, shows perfect LD with it. The minor allele of the SNP is always observed in the replacement-carrying haplotypes and the major allele always in the non-carriers.

## 4.6. Conclusion and Discussion

Results presented in this chapter have shown how the PRT-based measurement system for DEFA1A3 when used to type three-generation families of European origin has allowed inference of haplotype copy numbers in terms of three components: total copy number, copy number for the DEFA3 gene, and for the Indel-5 alleles. Segregation for this locus has been checked against the segregation information of surrounding genetic markers. Total haplotype copy numbers have been found from 1 to 7, a range that was expected given the measured diploid copy numbers in Europeans (3 to 11). Haplotype information obtained has also shown characteristics such as presence/absence of the DEFA3 gene with respect to Indel-5 minor allele copy numbers. In the 17 families studied, no instance of recombination in this locus has been observed, at least none that would detectably change copy numbers. A total of 338 transmissions have been observed, which gives a 95% confidence on the lower limit on the mutation rate to be  $8.86 \times 10^{-3}$  mutations per meiosis. In contrast, the CNV beta defensin locus on 8p23.1 has been shown to be highly mutable ( $\sim 0.7\%$  copy number changing mutations per meiosis) (Abu Bakar *et al.*, 2009).

Further analysis has been done by combining European copy number haplotypes with SNP genotypes (from the HapMap database) to search for potential copy number-tagging SNPs. This has shown several SNPs in good LD with copy number characteristics of the haplotypes, most notably rs4300027 that differentiates between low and high-copy haplotypes. This association has been verified by genotyping this SNP in an independent set of European samples. Non-European HapMap samples do not show associations between SNP genotypes and diploid copy numbers. However, the Japanese samples show other distinct copy number haplotypes that are tagged by surrounding SNPs. Some of them show a weak tagging ability in Europeans and Chinese. This seems to suggest that this locus has, due to a bottleneck effect and/or selective forces, acquired distinct features in the European and Japanese populations by the propagation of only a few haplotypes of the complete total present in the Chinese and African populations. This was already apparent in



the Japanese population from their measured diploid copy numbers, which go up to 16, while in the Chinese only up to 10. Thus high copy number haplotypes that may be very rare in the Chinese, have become frequent in the Japanese. This is why SNPs that differentiate high copy from low copy haplotypes in the Japanese have different allele frequencies between the Chinese and Japanese, with the high-copy haplotype associated allele being much more frequent in the Japanese. Another explanation for this observation could be that the two populations do not differ significantly, and the 45 samples from the Chinese population are non-random with respect to the *DEFA1A3* locus, such that high copy numbers have been inadvertently omitted and therefore also the SNP frequencies are shifted. However, as stated in Section 3.1, the probability of getting such a high difference based on chance alone is low (p-value=  $1.4 \times 10^{-14}$ ).

Thus, rs4300027 or any SNP in complete LD with it can be used as a proxy for *DEFA1A3* copy numbers in Europeans only, and can be genotyped in future studies for a broad classification of individuals into low, medium and high *DEFA1A3* copy number groups or even p-values for this SNP in already published GWAS can be investigated as has been done in the next part of this study (Chapter 5).

## 5. INTERROGATING GENOME WIDE ASSOCIATION STUDIES (GWAS)

Since the first GWAS using SNP microarrays in 2005 (Klein *et al.*, 2005) that targeted age-related macular degeneration, there have been hundreds more published to date (1,319 published GWAS till March 2011; [www.genome.gov/gwastudies](http://www.genome.gov/gwastudies)). These studies are largely SNP interrogating studies with 100,000 SNPs being the minimum cut-off for consideration in this catalogue. Large-scale SNP genotyping was made possible because of the tremendous increases in SNP databases resulting from information from the human genome sequence (2001, Venter *et al.*, 2001) and the SNP consortium (Sachidanandam *et al.*, 2001). The second important knowledge-base for designing and using SNP arrays came from definition of common haplotypes. This information came from the HapMap consortium (Gibbs *et al.*, 2003). Since there are only a finite and small number of common haplotypes in a given population, their knowledge allows using a minimal number of SNPs on the arrays that are sufficient to impute the genotypes for the excluded ones (Daly *et al.*, 2001). In the last decade, these databases have grown to include more and more SNPs, and several human populations (Altshuler *et al.*, 2010), resulting in increasingly better SNP arrays and thus improving the power and scope of GWAS. Ultimately, however, in light of economic constraints, the power of a SNP array to allow discovery is not necessarily directly proportional to the number of SNPs it genotypes but is affected by choosing the right set of SNPs dictated by population haplotypes and SNP allele frequencies (Spencer *et al.*, 2009). Most GWAS have been case-control studies while others are family-based designs. In case-control association studies, SNPs that show a significantly different genotype pattern between cases and controls are considered potentially associated with the phenotype in question. To get to such associated variants, the raw data is put through several quality control steps and statistical analyses to both impute excluded SNPs and remove genotyping artefacts. Also, because a large number of SNPs are genotyped in each study, and thus a large number of independent tests are made of the null hypothesis, the p-values need to be corrected for multiple testing, or in other words a much lower p-value is considered significant than would be for a single-locus study. Usually, p-values equal to or lower than  $10^{-7}$ - $10^{-8}$  are considered significant in GWAS. The reason why real associations are not always clear-cut is because these studies mostly deal with complex traits where each contributing locus has low penetrance and thus a

small effect size. Although care is taken to avoid getting false associations by matching individuals representative of the two pools as much as possible, because the number of SNPs typed is so large (>100,000) there still exists some chance for this error. Many GWAS results are verified by meta-analyses where data from different studies looking at the same trait may be combined, and top SNPs are genotyped in an independent set of samples, as was done for diabetes by the DIAGRAM consortium (Zeggini *et al.*, 2008). This allows refinement and increased power to detect associated variants.

## 5.1. SNPs as CNV Surrogates

Whether the associated SNPs found from GWAS are themselves causal or simply associated with causal variants is determined in the next stage. While considering CNVs as causal variants the question is how well do SNPs act as surrogate markers for them if at all. More specifically, when the interest lies in the dosage-effect of CNV rather than sequence variants in it, SNPs that differentiate between different copy number classes in a useful manner are the ones of interest. While SNPs that tag some variant within a CNV irrespective of the copy number haplotype, they could allow testing the effects of those variants rather than copy number. For simple CNVs, deletions and duplications with only two kinds of copy number alleles, there is strong LD with local SNPs (Hinds *et al.*, 2006). This is because they are also the same kind of bi-allelic variant that most SNPs are, where the variant has arisen only once and then spread in the population, with all examples of the derived state descended from a single ancestor. For multiallelic CNVs, there can be any number of different alleles resulting through multiple instances of copy number changes such that the same copy number allele could have arisen more than once on a different haplotypic background each time. However, whether there is strong LD or not depends upon the age, genomic structure and evolutionary history of the CNV in question. As has been shown earlier in this thesis (see Section 4.3), *DEFA1A3* CNV in the European population has surprisingly good LD with some surrounding SNPs. How many other such cases exist is unclear as there is a dearth of studies measuring multiallelic CNVs and examining their haplotypes, but this example is most likely to be an exception (Locke *et al.*, 2006). The CNVs involving beta-defensin genes (also on human chromosome 8) and *CCL3L1* on the other hand show a breakdown of LD with surrounding SNPs (unpublished work by R. Palla and S. Janyakhantikul) despite having lower levels of copy number variation than the alpha-defensin locus in Europeans. The discussion so far has excluded SNPs within the CNV regions. Such SNPs would be useless in bi-allelic deletion CNVs, but they may show

correlation with copy number haplotypes in simple duplications and multiallelic CNVs. They can also be more difficult to ascertain for SNP-status and genotype, especially in CNVs with higher copy numbers, and therefore do not pass the quality control to make it to SNP genotyping arrays.

## **5.2. rs4300027 as a Surrogate for *DEFA1A3* Copy Number**

As described earlier (Section 4.3), a regression analysis of the relationship between rs4300027 genotype and *DEFA1A3* diploid copy number gives a p-value of  $4.58 \times 10^{-28}$ , with the CC genotype associating with lower copy numbers and TT genotypes with higher copy numbers. The chi-squared p-value for the distribution of copy numbers in individuals with CC genotypes when copy numbers are classed into two groups (4 to 7 and 8 to 11) is  $3.8 \times 10^{-6}$ . For TT genotypes, the same analysis gives a p-value of  $3.3 \times 10^{-6}$ .

Heterozygotes (TC genotypes) were analysed by grouping copy numbers as one group of 4, 5, 9, 10 and 11 copy numbers and the other with 6, 7 and 8 copy numbers. The rationale for this partition is that if 2 and 3-copy haplotypes have allele C and 4 and 5-copy haplotypes have allele T, then heterozygotes would largely fall in the copy number range of 6 to 8 ( $2+4=6$ ,  $2+5=7$ ,  $3+4=7$ ,  $3+5=8$ ). Thus the observation of TC genotypes should be significantly different between the mid-range and tail end diploid copy numbers. This chi-squared test gives a p-value of  $1.5 \times 10^{-4}$ . Of the 159 ECACC samples typed in this analysis, one CC genotype had a copy number of 7, and two TT genotypes had a copy number of 6 and thus do not confirm to the low-versus-high categorization of these genotypes. Similarly, five TC genotype-individuals had a copy number of 4 or 5, five had a copy number of 9 and one had a copy number of 10 (Figure 4.8). While there may exist instances of incorrect copy number assignments to these samples, these deviations nonetheless should mostly come from the imperfect LD between this SNP and *DEFA1A3* copy number haplotypes. From the initial combination and comparison of HapMap individuals' haplotype copy numbers from segregation analysis with their SNP genotypes, it was observed that two 2-copy and two 3-copy haplotypes had allele T. Although none of the 4- and 5-copy haplotypes had allele C, this observation could be due to small sample size (25 haplotypes). The highest overlap of genotypes in any diploid copy number class is in 6-copy individuals, and then in 8-copy individuals. This is because a total of 6 copies can result from either two 3-copy haplotypes, or a combination of 2 and 4-copy haplotypes. In the first case, the genotype is likely to be CC, and in the latter case to be TC. Similarly, 8 copies can result from both 4+4 (TT) and 5+3 (TC). However, these overlaps do not affect so

much the power of predicting the copy number classes (low, medium, high) from a given genotype as they do predicting genotype from a given copy number. For example, if ECACC samples are divided according to their rs4300027 genotypes, then CC genotypes are 97% low copy number (4, 5 or 6), TC genotypes are 87% medium copy number (6, 7 or 8) and TT genotypes are 80% high copy numbers (8 or more). As explained earlier, 6-copy and 8-copy samples will unavoidably overlap two genotype classes.

### 5.3. Investigating GWAS

Hundreds of different human traits, including many diseases with an autoimmune component, have been studied through GWAS. The GWAS database allows one to search for the SNPs with significantly low p-values from all studies by entering the rs ID of a SNP or a different parameter. A search with rs4300027 or rs4512398 did not return any study. This was expected: if it had been a top SNP in a study, although the association with *DEFA1A3* copy number haplotypes would have been unknown, the physical proximity of it to the *DEFA1A3* repeat region (about 800 bp from the telomeric end) would have led researchers to label these genes as potential candidates for affecting the phenotype in question. Previously searching the database with the gene name (DEFA1/DEFA3/HNP1/HNP3) had returned nothing.

For investigating whether this SNP showed a significantly different distribution between cases and controls of a disease, much higher p-values, in the range of  $10^{-3}$ , can be considered significant. This is because the issue of multiple-testing involved in a GWAS does not exist, because we are now investigating one SNP and not >100,000. This meant that p-values for rs4300027 had to be looked up in GWAS of interest. However, due to concerns over protecting the identity of individuals participating in these studies and their genotypes (Homer *et al.*, 2008), all summary data from GWAS apart from the top hit SNPs is not made public anymore. Therefore, this investigation involved drawing up a list of GWAS and contacting relevant authors of the studies with a request to share the p-value observed for rs4300027 or rs4512398, whichever was represented on the genotyping array used.

Given the functional role of defensins as antimicrobial and immunomodulatory peptides, the most likely effect of dosage variation would be in autoimmune and inflammatory diseases, and those were of primary interest. The following diseases were investigated:

**Coeliac disease** results from an autoimmune destruction of the epithelial lining of the small intestine, which is triggered by proteins in wheat and other

cereals (Alaedini *et al.*, 2005). The main immune components involved are T-cells, and their stimulation by antigen-presenting HLA molecules. A strong genetic predisposition in individuals carrying certain HLA alleles has been found. GWA studies looking for further genetic factors predisposing to this disease have been carried out (Dubois *et al.*, 2010). This study has identified risk loci in genes involved in T-cell development, innate immune detection, stimulation of B- and T-cells and production of cytokines and their receptors. Although this disease appears to be a largely lymphocyte-based reaction, the role of neutrophils in inflammation and influencing the adaptive immune response may still allow variation in them or their components to affect disease development (Diosdado *et al.*, 2007). The GWAS by Dubois *et al.* studied 4,533 cases (2,586 UK, 647 Finland, 803 The Netherlands and 497 Italy) and 10,750 controls (7,532 UK, 1,829 Finland, 846 The Netherlands and 543 Italy), and should be well-powered to detect such further risk variants.

**Inflammatory Bowel Disease** (IBD) is a group of chronic inflammatory conditions of the gastrointestinal tract, the two major ones being **Crohn's disease** and **Ulcerative Colitis** (Cho, 2008). Both diseases are characterized by chronic inflammation of the intestinal epithelia and sometimes one is misdiagnosed for the other, or they are difficult to tell apart. However, there are distinct differences in the histopathology and location of the two which allows their characterization (Cho, 2008). From association studies, some genetic loci have been found to predispose to one and not the other, and some are common between the two. For example, a variant of an interleukin gene, *IL23R*, is strongly associated with both, and a variant of *NOD2*, a receptor for pathogen molecules, is strongly associated with Crohn's disease but not with Ulcerative Colitis. IBDs are thought to result from an abnormal response to the gut microbes. Neutrophil infiltration is present in inflamed mucosa and thus varying neutrophil defensins may affect disease development. Three IBD GWAS have been investigated: one has studied 3,230 cases of Crohn's disease and 4,829 controls of European ancestry (Barrett *et al.*, 2008), second has studied 1,052 cases of Ulcerative Colitis and 2,571 controls of European ancestry (Silverberg *et al.*, 2009) and the third has studied 1,011 cases of IBD and 4,250 controls of European ancestry (Kugathasan *et al.*, 2008).

**Psoriasis** is another chronic inflammatory disease that affects the skin. It is also considered a T-cell mediated autoimmune disease with infiltration of various leukocytes and a complex underlying mechanism (Bowcock *et al.*, 2005). Beta-defensins produced by the skin epithelia are thought to contribute to disease predisposition and their variable gene copy number has been found associated with psoriasis risk (Hollox *et al.*, 2005). A role for

neutrophil defensins was investigated in data from two GWA studies via rs4300027 association: one studied 318 psoriasis cases and 288 controls of Northern European descent (Capon *et al.*, 2008) and the other studied 223 psoriasis cases, including 92 cases of psoriatic arthritis, and 519 controls of European descent (Liu *et al.*, 2008). About 25% of psoriasis patients also develop psoriatic arthritis which has no correlation with psoriasis severity.

**Type-1 Diabetes** results from an autoimmune destruction of the insulin-producing cells of the pancreas. It is the T-cells that destroy the pancreas, and their recruitment involves cells and signalling molecules of the innate immune system (Lehuen *et al.*, 2010). The GWAS that was interrogated studied 563 patients, 1,146 controls and 483 family trios, all of European descent (Hakonarson *et al.*, 2007). This study found an association with a gene coding for a sugar-binding lectin. Previous studies have found associations with the insulin locus, HLA locus, T-lymphocyte associated protein 4 and an interleukin amongst others.

**Multiple Sclerosis** is another autoimmune disease, in which the immune system destroys neural cells in the central nervous system. It also has a complex mechanism, starting with the breakdown of the blood-brain barrier, resulting in lesions in the brain which are visible on brain scans, and yet it is not clearly understood how these processes occur (McFarland *et al.*, 2007). The GWAS that was interrogated studied 931 family trios (affected child and both parents, 453 from the UK and 478 from the USA) and 2,431 controls (Hafler *et al.*, 2007). That study identified risk variants in interleukin and HLA genes.

**Atopic Dermatitis** is also a chronic and recurrent inflammation of the skin resulting in lesions that are pruritic and have reduced defence against microbes (unlike psoriatic lesions, which are highly antimicrobial). The skin of atopic dermatitis patients, even without lesions, has a higher microbial burden than normal (De Benedetto *et al.*, 2009). The observed anomalies in this disease are a breakdown of the skin barrier, decreased innate immune response and enhanced allergic response. Neutrophils are absent from lesions despite infection. The GWAS that was interrogated studied 1,468 patients and 1,543 controls, all of European descent (Gordillo *et al.*, 2009). The only variant shown to affect predisposition to atopic dermatitis is loss-of-function mutations in the gene coding for filaggrin, which is a component of the skin barrier.

Table 5.1 lists the p-values obtained for rs4300027/rs4512398 or an alternative SNP, rs7825750, in cases where rs4300027 was not genotyped, in GWAS studying various immunity-related diseases.

**Table 5.1** P-values obtained for *DEFA1A3* copy number-tagging SNPs in GWAS studying autoimmune/inflammatory phenotypes and corrected p-values for multiple testing

Phenotype	SNP	p-Value**	Corrected
Celiac Disease	rs4512398	0.013	0.105067161
Crohn's Disease	rs4512398	0.525	0.997443079
Psoriasis	rs4512398	0.2	0.83222784
Psoriasis	rs4300027	0.49	0.995915281
	rs4512398	0.926	0.999999999
Ulcerative Colitis	rs4512398	0.8	0.99999744
	rs4300027	0.71	0.999949975
IBD and T1D*	rs4512398	>0.05	
	rs4300027	>0.05	
Multiple Sclerosis	rs7825750	0.5904	0.999207718
Atopic Dermatitis	rs7825750	Not significant	

\*IBD= Inflammatory Bowel Disease, T1D= Type-1 Diabetes

\*\* We are grateful to the following people for sharing their GWAS data: David van Heel, Michel Georges, Richard Trembath, Anne M. Bowcock, Richard H. Duerr, Hakon Hakonarson, Young Ae-Lee and Stephen Sawcer.

As the table shows, none of the studies have significant or even suggestive p-values for rs4300027. The lowest uncorrected p-value is 0.013 in the celiac disease GWAS, which is ten times higher than the  $10^{-3}$  range considered significant in such a large study. In some genotyping platforms, neither rs4300027 nor rs4512398 was included. In those cases the next best available SNP was rs7825750, which gave a correlation coefficient of 0.3 with haplotype copy numbers divided into low (2 and 3) and high (4 and 5) (rs4300027 gave a value of 0.67). Since p-values (either for rs4300027 or rs7825750) from eight



independent GWAS were interrogated, this increases the probability of finding an association purely by chance than if only one GWAS was investigated. Thus, each p-value is corrected for multiple testing, i.e. of testing several studies, by applying the formula  $1-(1-p)^n$ , where p denotes the p-value reported in the GWAS, and n is the number of GWAS investigated.

## 5.4. Conclusions and Discussion

SNP rs4300027 is a good proxy for *DEFA1A3* copy number in Europeans, as evidenced by percent accuracy of predicting diploid copy number class (low, medium and high) from genotypes. This means that for large-scale studies in the European population *DEFA1A3* copy number can be effectively measured by genotyping this SNP and avoiding the more cumbersome process of genotyping the CNV itself, as long as broad copy number class is a good enough measure. This useful property has been exploited to investigate GWAS which have studied phenotypes where neutrophil alpha defensins might play a role. These GWAS are well-powered with sample sizes ranging in the thousands. In the seven disease phenotypes investigated here, none has returned a significant p-value for rs4300027 (or SNPs in LD with it). As discussed earlier (Section 5.1), other multiallelic CNVs that have been studied by others do not show linkage with surrounding markers in a way that would allow using them as surrogates for copy number.

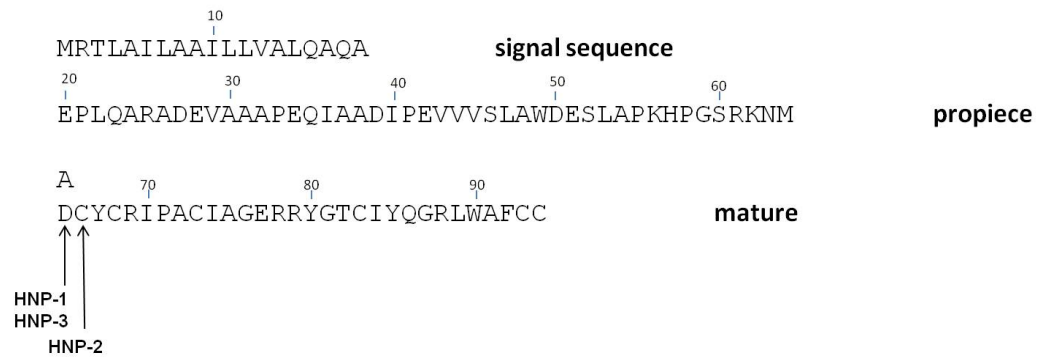
Here an assumption has been made, implicit so far, that gene copy numbers of *DEFA1A3* will directly affect protein levels such that higher the copy number, greater the protein expression level. In reality, it is unknown how expression levels are affected by this CNV. The next chapter details a methodology for quantifying gene expression of *DEFA1A3*, and this is an instance where the exact copy number of each sample tested is of interest, rather than which of the three broad copy number classes it belongs to.

## 6. EXPRESSION OF *DEFA1A3* GENES

### 6.1. Synthesis of *DEFA1/2/3* peptides

Transcription and translation of *DEFA1* and *DEFA3* genes occurs in the neutrophil precursors, the promyelocytes in the bone marrow, and is not normally observed in other body tissues (Daher *et al.*, 1988). The exceptions found by Daher *et al.* are tissues where neutrophils or their precursors are present (for example, peripheral blood in leukemia patients) and eosinophils have also been shown to produce alpha defensins (Driss *et al.*, 2009), thus restricting the production of these peptides to the myeloid lineage of blood cells. However, neutrophils are the largest source both because defensins make up almost 5% of total neutrophil protein content (1 to 5  $\mu\text{g}$  per  $10^6$  neutrophils) (Ganz, 1987, Harwig *et al.*, 1992) and because neutrophils are the most abundant white blood cells (50% to 70% of total white blood cells), whereas the other granulocytes, basophils and eosinophils are the least frequent (0.5% to 3% of total white blood cells) in normal individuals. Further evidence of myeloid-specific expression of these genes comes from studies of their transcriptional control. It has been shown that a CCAAT/enhancer-binding protein (C/EBP $\alpha$ ) positively regulates the transcription of *DEFA1* and *DEFA3*, and does so equally well for both (Tsutsumi-Ishii *et al.*, 2000). Experimental removal of C/EBP binding site in the defensin genes' promoter region does not abolish transcription but reduces it to less than half. However, in Specific Granule Deficiency (SGD) caused by loss-of-function mutations in C/EBP $\epsilon$  gene, besides absence of secondary and tertiary granule proteins in neutrophils, alpha defensins are also reduced (Gombart *et al.*, 2001, Lekstrom-Himes *et al.*, 1999). The link between C/EBP $\alpha$  and C/EBP $\epsilon$  is most probably that C/EBP $\alpha$  controls the expression of C/EBP $\epsilon$  which has been shown to be specific to the myeloid/lymphoid progenitors, and C/EBP $\alpha$  has been shown to act earlier in the development of these lineages (Lekstrom-Himes *et al.*, 1998).

The mRNA's for these defensins (as for others) code for a pre-prodefensin peptide that is 95 amino acids long and includes a signal peptide at the amino terminal, followed by a propiece and then the mature peptide (Valore *et al.*, 1992). This pre-prodefensin is processed to a 75 amino acid long prodefensin by the cleavage of the signal peptide sequence. This is then further processed to a 56 amino acid prodefensin, which is cleaved to give the 30 amino acid *DEFA1* and *DEFA3*, and the 29 amino acid *DEFA2* mature peptides. Valore *et al.* found this processing to take 4 to 24 hours in normal and leukemic cultured cells. Harwig *et al.* also report other intermediate prodefensin peptides of 39, 34 and 32 amino acids in the neutrophils but in minute amounts (0.25% of total defensin content) (Harwig *et al.*, 1992).



**Figure 6.1** Amino acid sequence of DEFA1/2/3 (HNP-1/2/3) (Valore *et al.*, 1992)

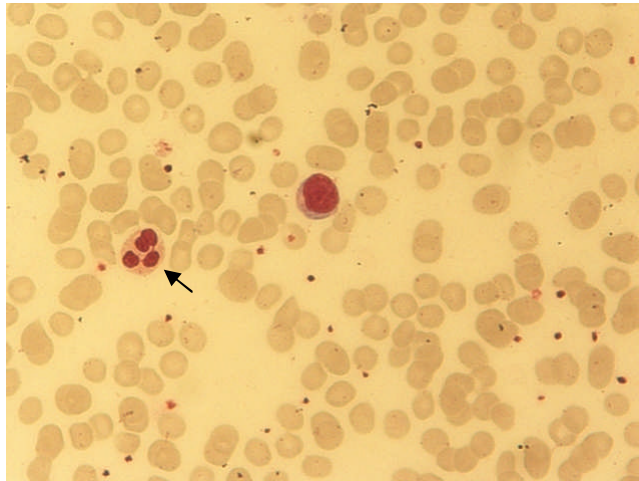
The defensins are packaged in the azurophil (primary) granules of the neutrophils. The propiece is necessary for the targeting of the peptide to these granules and processing to the active form takes place post-packaging (Gullberg *et al.*, 1999). Also, the development of the three main granules of neutrophils (azurophil, specific and gelatinase) occurs in a temporally stratified sequence and the timing of the production of granule proteins influences which granules they are sorted to (LeCabec *et al.*, 1996).

Before any concrete knowledge of the highly polymorphic CNV of *DEFA1A3*, the lower copy numbers of *DEFA3* as compared to *DEFA1*, and the absence of the *DEFA3* gene in some proportion of the human population was available, it was already observed that the DEFA1 peptide accounted for a larger proportion of total neutrophil defensin content than the DEFA3 peptide (Chertov *et al.*, 1996). With a much better characterization of this CNV, how it affects expression levels is an interesting and obvious next question. As shown by Tsutsumi-Ishii *et al.*, the promoter and enhancer (C/EBP binding site) are within 200 bp upstream of the first exon of the gene. However, there could be several other factors influencing gene expression in the complex CNV arrangement and only a direct measurement of gene copy numbers and of gene expression will address this question. The measurement of genes has already been discussed in the previous chapters. For gene expression measurements, neutrophils are an easily obtained cell type from the blood as opposed to obtaining promyelocytes from the bone marrow. This exclusion of promyelocytes means that gene expression has to be studied at the protein level rather than mRNA level. Aldred *et al.* used neutrophils for defensin mRNA quantification and found no correlation with total gene copy numbers (Aldred *et al.*, 2005). They did find that *DEFA1:DEFA3* gene ratio correlated with DEFA1:DEFA3 mRNA ratio, and acknowledged that protein measurements from neutrophils would give a more accurate quantification of expression.

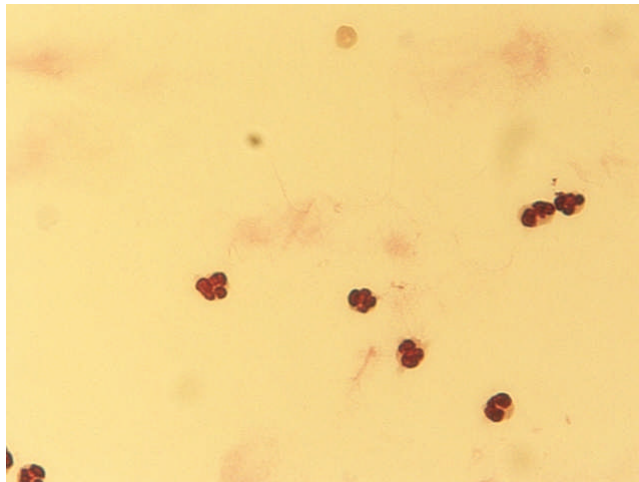
## 6.2. Neutrophil variability

Neutrophils are the most abundant of the white blood cells (60% to 70%) and also very variable in number from person to person. Some variation in neutrophil counts depends upon age, sex, ethnicity and even time of day, but even within people matched for these factors, there is great variability.

Further variation comes from infection or other pathophysiological differences which may not be apparent or measurable, and hence cannot be accounted for. This high variability is possible because of their high turnover rate: neutrophils survive for at most 5 days in the blood (Pillay *et al.*, 2010). In healthy, European adults (median age = 25 years) neutrophil counts range from 1500 to 6500 million per litre of blood (Greer, 2009). This means that rather than taking all neutrophils from the same volume of blood as a starting source of defensins, neutrophils need to be counted for each blood sample (even when repeating for the same individual). Section 13 of Chapter 2 describes the methods used to obtain granulocytes (neutrophils, eosinophils and basophils) and count them. To demonstrate the separation of granulocytes from other blood cells, Figure 6.2 and Figure 6.3 show micrographs of stained slides of whole blood and separated granulocytes. About 97% of the separated cells were granulocytes which is the expected frequency from the protocol used.



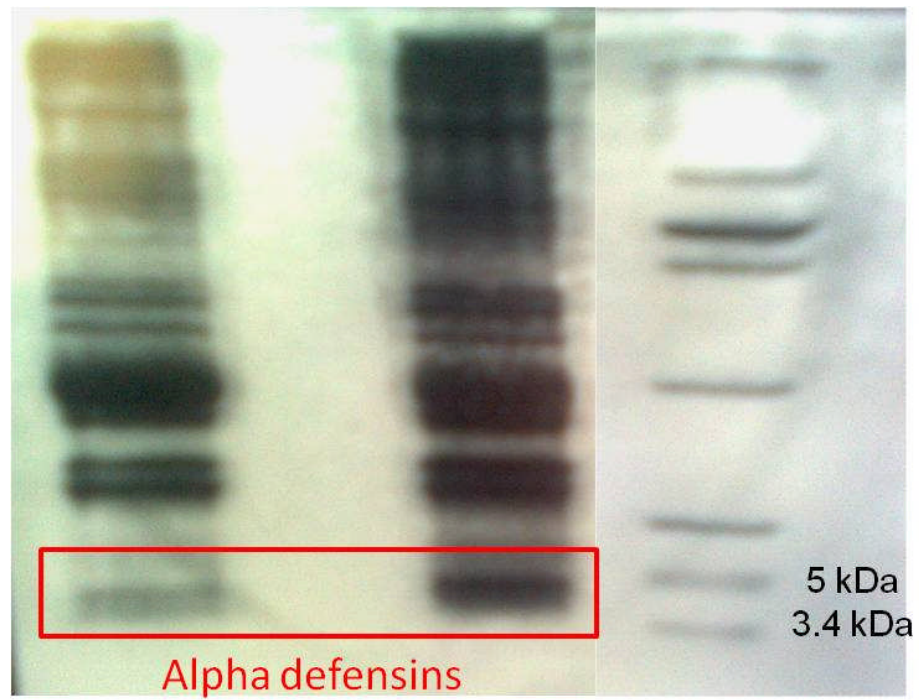
**Figure 6.2** Micrograph of Leishman-stained whole blood smear (400x). Arrow points to a neutrophil.



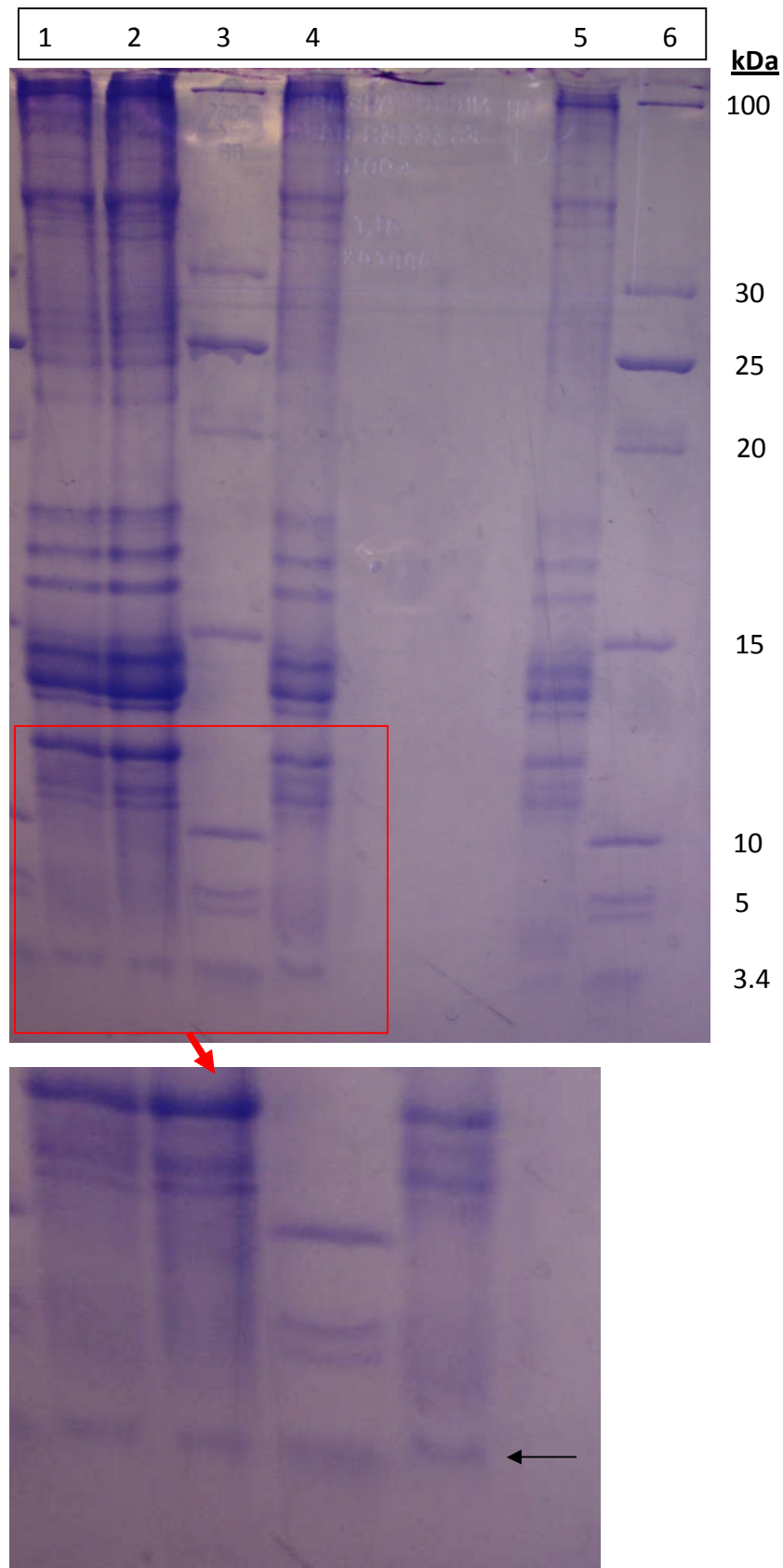
**Figure 6.3** Micrograph of Leishman-stained separated granulocyte smear (400x). Cells have lost some volume during the processing and appear smaller than in whole blood.

### **6.3. Determining Alpha Defensin Peptides in Neutrophil Extracts**

Separated neutrophils were treated with acetic acid (as described in Section 2.13) to extract proteins, including granule proteins. To confirm the presence of defensins in the neutrophil extracts, the extracts were subjected to 18% Tris-tricine SDS PAGE. The molecular weight of neutrophil defensins (DEFA1, 2, 3 and 4) is 3.5 to 3.9 kDa. Since most neutrophil proteins are bigger than this size and since defensins are present in large quantities, the band in the gel corresponding to 3.5 to 4 kDa size would contain only or mostly the defensin peptides as has been demonstrated previously (Ganz *et al.*, 1985). This band was cut out and sent for sequencing to another laboratory (Figure 6.4). Note that since a small gel (8 x 10 cm) was used, the resolution is poorer than obtained by Ganz *et al* on a longer gel. Better resolution was obtained in later experiments using large gels (18 x 20 cm) (Figure 6.5). Note that the well separated defensin band in the large gel is not of an intensity expected of a peptide that makes up 5% of neutrophil proteins, and that is due to poor fixation of small sized peptides in gels after electrophoresis. The band containing the defensin peptides in the smaller gel is more intense probably because it contains other peptides as well due to poorer resolution.



**Figure 6.4** Photograph of a Coomassie Blue-stained SDS PAGE gel (18% Tris-tricine) with two samples of granulocyte extracts from  $10^6$  cells each and a protein ladder in the right most side. The bands from the cell extracts that correspond to the size of alpha defensins and that were sequenced are highlighted in the red box



**Figure 6.5** Photograph of a Coomassie blue-stained Tris-tricine SDS PAGE (18%) with magnified image of the gel area with the defensin bands (indicated with black arrow) below. Lanes 3 and 6 are protein ladders, lanes 1, 4 and 5 are extracts from  $5 \times 10^5$  neutrophils and lane 2 is extracts from  $10^6$  neutrophils.

The sequences of the neutrophil peptides are:

(DEFA1)	ACYCRIPACIAGERRYGTCTIYQGRLWAFCC
(DEFA3)	DCYCRIPACIAGERRYGTCTIYQGRLWAFCC
(DEFA2)	CYCRIPACIAGERRYGTCTIYQGRLWAFCC
(DEFA4)	VCSCRLVFCRRTLRVGNCLIGGVSTTYCCTRVD

However, the sequencing method used was not able to determine Cysteines (Edman degradation method) and only the first six residues were sequenced. Additionally, the sample sent for sequencing came from a person without the *DEFA3* gene (determined by the DefHae3 assay, Section 2.2) which meant that the sequence expected was a mixture of the following three:

(DEFA1)	A-Y-RI
(DEFA2)	-Y-RIP
(DEFA4)	V-S-RL

The sequencing results obtained are shown in Figure 6.6. The first amino acids are as expected Ala and Val, the second includes Tyr (from DEFA2), the third is Tyr (from DEFA1) and Ser (from DEFA4), the fourth includes Arg (from DEFA2), the fifth should have been predominantly Arg (from DEFA1 and DEFA4) but is only weakly detected for some reason and Ile (from DEFA2), the sixth is as expected predominantly Ile (from DEFA1), but no Pro (from DEFA2) or Leu (from DEFA4) are detected. It was indicated that Pro residues are often not recognized (personal communication with the sequencing lab). There are clearly some other sequences mixed with the defensins which give a His at position 2, Asn at 3, Lys at 4 and Ala at 5. These could come from other similar sized proteins or proteolytic fragments from larger proteins in the neutrophils. Since these could be a complex mixture and defensins were identifiable despite their presence, no attempt was made to identify the contaminants.



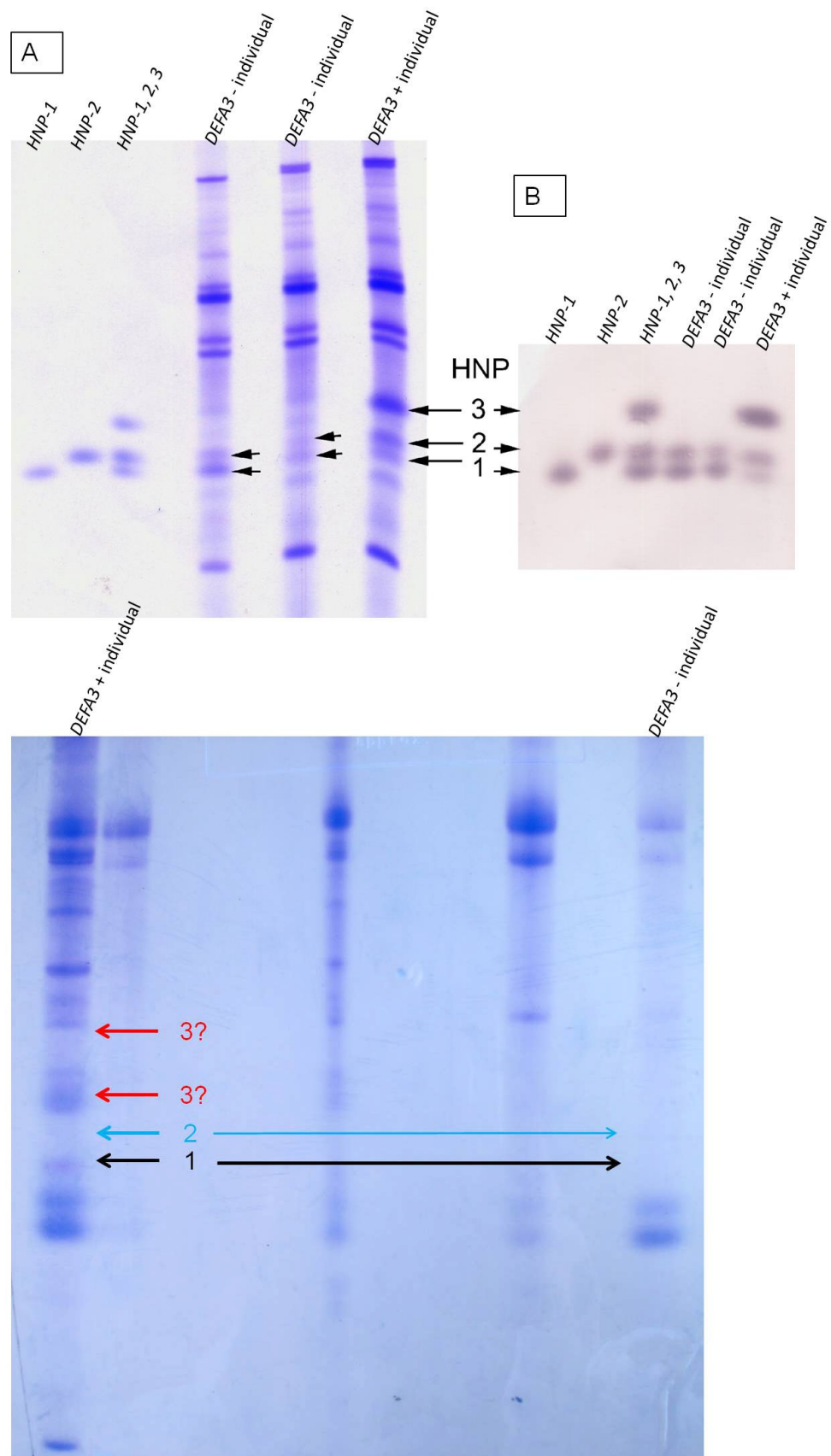
### N terminus

Residue			
1	A	V	
2	H	Y	L
3	Y	N	S
4	L	R	P?
5	A	I?	R?
6	I	D	

**Figure 6.6** The result of peptide sequencing. Single letters represent amino acids, with those in the left most column being the predominant ones at the corresponding position and those in the right most column being the least concentrated ones.

## 6.4. Acid Urea PAGE

Acid urea Polyacrylamide gel electrophoresis (AU PAGE) was carried out for neutrophil extracts as described in Section 2.13. The purpose of this electrophoresis system is to allow separation of very similar or same-sized proteins, such as DEFA1, 2 and 3 on the basis of charge differences due to differences in amino acid composition. At the low pH of the gel (about 3) even amino acids with the lowest  $pK_a$  are protonated, resulting in a positive charge on all proteins and therefore migration is from anode to cathode (as opposed to cathode to anode in an SDS PAGE system). Separation occurs on the basis of both mass and strength of the positive charge on the proteins. In this experiment no standard defensin peptides for reference were included in the gel, nor were the separated peptides confirmed by a Western blot or any other means. Since separation on an AU PAGE is based on both size and charge of the proteins, there are no standard markers available. Thus the only means of hypothesizing which of the bands correspond to the defensins was by comparing the obtained profile with that obtained by previous studies, Linzmeier and Ganz in this case (Linzmeier *et al.*, 2005) (Figure 6.7). The band containing the DEFA3 peptide could be either of the two bands marked with red arrows, as the *DEFA3*- individual in lane 2 has several bands missing in the profile and not just the DEFA3 band as it ideally should have.

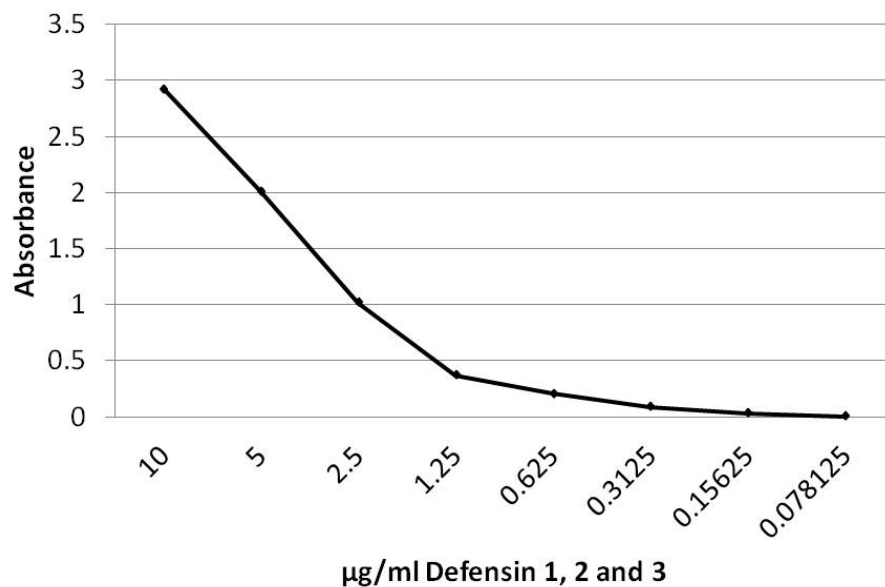


**Figure 6.7** The top part of the figure has been taken from Linzmeier et al (Linzmeier *et al.*, 2005), where part A shows a Coomassie blue-stained AU PAGE where the first three lanes

contain standard defensin peptides, lanes 4 and 5 contain neutrophil extracts from individual(s) lacking *DEFA3*, and lane 6 contains neutrophil extract from an individual possessing *DEFA3*. Part B shows the western blot. The gel at the bottom is from this study, where the left most lane contains neutrophil extract from a *DEFA3*+ individual and the right most lane contains neutrophil extract from a *DEFA3*- individual. The band corresponding to *DEFA3* is more likely to be the upper band marked with a red arrow, as it corresponds better with the position of bands with respect to Linzmeier et al's AU PAGE.

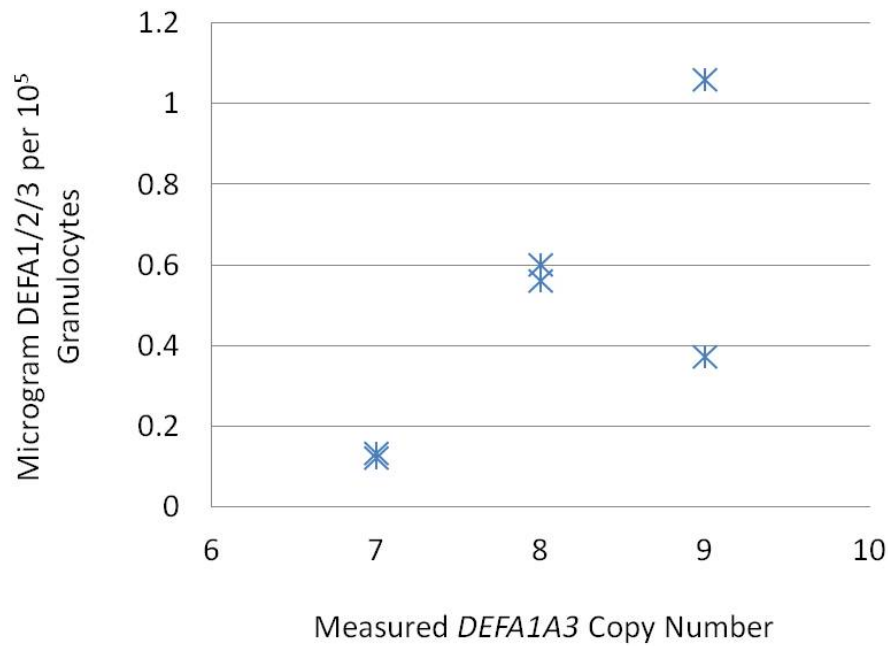
## 6.5. Quantifying Alpha Defensin Peptides (DEFA1/2/3)

For quantification of the alpha defensins peptides, ELISA was used (see Section 2.13). Three samples were typed in duplicate using purified defensins as standards. The standards ranged in concentration from 0.07 to 10  $\mu\text{g/ml}$ . Their absorbance values (at 450 nm) recorded at the end of the ELISA are plotted against their concentrations in Figure 6.8.



**Figure 6.8** Graph showing absorbance values obtained from the ELISA of defensin standard proteins

Given the defensin amounts estimated previously, concentrations of  $10^5$  granulocytes per ml were used for analysis. Three samples were tested in duplicate, i.e. two extracts from the same sample were tested in the same experiment. The absorbance recorded was used to calculate defensin amounts using the standard curve (Figure 6.8) by non-linear regression equation in GraphPad Prism software. These calculated amounts are plotted against the measured *DEFA1A3* gene copy numbers for the same samples (Figure 6.9).



**Figure 6.9** Calculated defensin amounts from ELISA end-point absorbance values plotted against measured gene copy numbers.

The standard deviation between the duplicates of two samples (copy numbers 7 and 8) are low (0.009 and 0.028, respectively), but for the third sample is quite high (0.485) as is evident from the data points on the graph. It is not possible to infer any relationship between gene copy numbers and protein amounts from this data as it has only three samples out of which one is highly inconsistent.

## 6.6. Investigating a Role for Alpha Defensins in Cystic Fibrosis

Cystic Fibrosis (CF) is an autosomal recessive disease that is characterized by pancreatic malfunction due to its fibrosis, and fatally abnormal lung function due to clogging of airways with mucus, resultant chronic infections and fibrosis of lung tissue (Davis, 2006). The current median age of survival is between 30 and 40 years (Davis, 2006). The genetic cause is a loss-of-function or decrease-of-function mutation in the Cystic Fibrosis Transmembrane conductance Regulator (CFTR) gene (Riordan *et al.*, 1989). This gene codes for a membrane channel that is responsible for active transport of chloride ions across the cell membrane. Mutations in *CFTR* lead to decreased or absent activity of this channel and when a person inherits two mutated copies of the gene, it can result in CF. This membrane channel is expressed in several organs of the body, mainly in cells of the epithelial tissue that line the airways, intestines, and exocrine ducts. These are all sites of production of fluids: mucus, digestive fluids or other exocrine secretions like

sweat and saliva. A sub-optimal function of pumping out chloride ions results in a change of osmotic pressure across the apical surfaces of these cells that causes more water to enter the cells and results in thickening of the extracellular fluids.

## CF Genetics

The *CFTR* gene is about 203 kb and the protein is 1,480 amino acids long. As of December 2011 the number of *CFTR* gene mutations catalogued in the database is 1,897 ([www.genet.sickkids.on.ca/StatisticsPage.html](http://www.genet.sickkids.on.ca/StatisticsPage.html)). These also include mutations in regulatory regions of the gene. Not all of these mutations result in CF, though they may cause or predispose to other pathologies, which also co-occur with CF. These include pancreatitis and congenital bilateral absence of vas deferens amongst other symptoms. Whether the two mutations inherited by a person will cause CF depends upon the nature of the mutations and how much *CFTR* function is affected (Griesenbach *et al.*, 1999). CF results when *CFTR* function is less than 10% of normal. Pancreatic insufficiency occurs only at a lower residual *CFTR* function of about 5% or less. Of all these mutations,  $\Delta F508$  (deletion of phenylalanine at position 508) is the most common: it has been found to have a relative frequency of 0.67 among disease-causing mutations in a combined cohort of European, north African and west Asian populations (Lao *et al.*, 2003). However, the frequency varies between these populations in a clinal fashion: it is highest in northern Europeans, followed by southern Europeans, west Asians (Turkey, Lebanon and Israel) and least in Tunisians. Since Europeans have the highest carrier frequency of CF mutations (1 in 25 people in the UK is a carrier) and  $\Delta F508$  predominates amongst them all, further discussion will pertain to this mutation and patients homozygous for it. It is rare to observe homozygotes for other *CFTR* mutations, and studies of those cases are limited (Stanke *et al.*, 2008).  $\Delta F508$  results in a *CFTR* product that is abnormally processed (Sharma *et al.*, 2004). It has been shown that the protein is retained in the endoplasmic reticulum and >99% is degraded (as opposed to 75% of normal protein). Furthermore, if this mutant protein reaches the apical membrane of the cell it is abnormally internalized and marked for degradation.

## Why Study CF?

Although  $\Delta F508$  homozygotes have CF, the lung disease phenotype is highly variable showing low correlation to *CFTR* genotype as opposed to the high correlation between pancreatic disease and *CFTR* genotype (Hamosh *et al.*, 1993), but it does show a genetic contribution to this variability rather

than only an environmental contribution (Vanscoy *et al.*, 2007). Since CF is the most common severe genetic disease in Europe, and the major cause of mortality is lung disease, it is of great interest to find other genetic factors that modify disease severity. Given the lung pathology observed in CF patients and discussed under the following heading, alpha defensins become an interesting potential modifier locus.

## CF Lung Disease

Lung disease may be the major killer in CF but despite tremendous research to understand its development it remains the least understood of all CF manifestations (Boucher, 2004). CF patients are born with normal lungs. The basic defect of CF results in abnormal airway surface liquid (ASL) volume and tonicity that leads to impaired clearance of the airways by adversely affecting the normal ciliary movement and the lung's defence mechanisms. This allows microbes that are not pathogenic in normal lungs to infect CF airways and establish chronic infections, which the immune system cannot clear. How the innate immune system of the lungs is affected by the basic defect of CF remains poorly understood. It was initially thought that the ASL in CF patients has a higher salt content than normal which inhibits antimicrobial peptides like the beta defensins produced by the lung epithelia (Goldman *et al.*, 1997). However this hypothesis has been challenged by studies showing that there is no difference in the tonicity of ASL between CF patients and normal controls (Knowles *et al.*, 1997). An alternative hypothesis implicates a decreased volume of ASL in impaired airway clearance and immune function (Matsui *et al.*, 1998). In either case, it is not disputed that the innate immune function of the airways is impaired, leading to infections by opportunistic pathogens. An almost characteristic chronic infection in CF lungs (especially in older patients) is that of *Pseudomonas aeruginosa* (Brugha *et al.*, 2011). This bacterium is easily cleared from normal airways without causing any inflammation. Besides the higher-than-normal susceptibility to infections, inflammation is also considered to play a role in the deterioration of the CF airways. Not only this, but inflammation and increased pro-inflammatory cytokines and cells are seen to occur prior to any infections (Khan *et al.*, 1995, Tirouvanziam *et al.*, 2002). This increased inflammation is thought to occur because of abnormal neutrophil recruitment and to result in epithelial damage and facilitation of microbial adherence (Plotkowski *et al.*, 1989). In contrast, other studies have found no change in pro-inflammatory cytokine concentrations or other evidence for inflammation prior to infection in CF lungs (Aldallal *et al.*, 2002, Muhlebach *et al.*, 2004).

## Components of Airway Innate Immunity

The airways are constantly exposed to microbes in inhaled air. Mechanical clearance consists of the mucus lining the epithelia that functions to trap particulate matter which is cleared out via the beating of the cilia of the epithelial cells (mucociliary clearance) (Houtmeyers *et al.*, 1999). Chemical immunity includes pathogen recognition molecules (Toll-like receptors, for example) on the surface of epithelial cells and secretion of antimicrobial substances, and inflammatory and chemotactic mediators (Bals *et al.*, 2004). These not only directly eliminate bacteria but also recruit neutrophils in case of establishment of infection (Craig *et al.*, 2009). While insufficient neutrophil infiltration can result in persistent infection, a high neutrophil accumulation can cause injury to lung tissue by increased and/or prolonged inflammation. Neutrophils are present in large numbers in the lung vasculature as compared to the bigger blood vessels because of their slower passage (Doerschuk *et al.*, 1993). This facilitates their quick recruitment into the airways.

## CF Lung Disease and Modifier Loci

Several studies have been undertaken to find genetic loci that act as modifiers of lung disease severity. Earlier studies focused on immunity-related genes and more recently genome-wide association studies have been performed. While some loci have given conflicting results, some have shown clear association or lack of it. However, the variability remains largely unexplained. Lung function is the standard diagnostic and prognostic measure of lung disease in CF (Ranganathan *et al.*, 2008, van der Ent *et al.*, 1999), and is used as the variable phenotype in most association studies. Lung function is measured by spirometry in terms of Forced Expiratory Volume in one second (FEV<sub>1</sub>) and Forced Vital Capacity (FVC) in case of adults and children aged 6 years or older. In younger children, more complicated techniques are adopted. Children under 6 years of age are usually not included in association studies.

Anti-proteases, like  $\alpha_1$ -antitrypsin ( $\alpha_1$ -AT), are present normally in the ASL and balance the presence of proteases, like neutrophil elastase, metalloproteases and others (Garcia-Verdugo *et al.*, 2010). In CF, increased proteases have been detected in the lungs (Birrer *et al.*, 1994). Studies have looked at  $\alpha_1$ -AT deficiency and CF lung disease and found conflicting results: increased susceptibility to *P.aeruginosa* infection but no adverse effect on lung function (Doring *et al.*, 1994), beneficial effect on lung function (Mahadeva *et al.*, 1998), and no effect on lung function (Frangolias *et al.*, 2003). Although these three studies included  $\Delta F508$  homozygotes,

heterozygotes and non- $\Delta F508$  CF patients, the effect of  $\alpha_1$ -AT haplotypes was also studied in  $\Delta F508$  homozygotes alone.

Tissue Growth Factor- $\beta$  (TGF- $\beta$ ) is a cytokine that can promote fibrosis of tissue and affect inflammatory response (Sporn *et al.*, 1986). A high-expressing TGF- $\beta$  haplotype has been found associated with a faster decline in lung function in CF (Arkwright *et al.*, 2000). Another study looking at 16 polymorphisms previously reported to be associated with CF lung function found TGF- $\beta$  haplotype as the only significant association in >800  $\Delta F508$  homozygotes (Drumm *et al.*, 2005). Other loci tested in this study included Tumor Necrosis Factor- $\alpha$  (TNF- $\alpha$ ), Mannose-binding Lectin (MBL), Interleukin-10 (IL-10), Nitric Oxide Synthase 3 (NOS3),  $\alpha_1$ -antiprotease ( $\alpha_1$ -AP) and Angiotensin Converting Enzyme (ACE) amongst others. These TGF- $\beta$  haplotypes associated with CF have also been found to affect disease progression in idiopathic pulmonary fibrosis (Xaubet *et al.*, 2003).

A recent GWAS using a SNP platform of 2.3 million SNPs has found two loci in significant association with lung disease severity in CF patients (1,978  $\Delta F508$  homozygotes and replicated in 557  $\Delta F508$  homozygotes) (Wright *et al.*, 2011). The nearest genes to these loci (SNPs) are ones coding for epithelial-specific transcription factors, a protein involved in inhibiting apoptosis, a neuromodulatory protein and a melanocortin receptor. They found p-values for TGF- $\beta$  in the range of  $10^{-3}$  to  $10^{-4}$  dependent upon covariates, but higher values on its own at the genome-wide level.

Beta-defensins are produced by the lung epithelia. A study of the beta-defensin CNV has shown no significant modifying effect on CF lung disease severity (Hollox *et al.*, 2005). This study looked at 355 patients, of which about 45% were  $\Delta F508$  homozygotes, 40% were  $\Delta F508$  heterozygotes and the rest had non- $\Delta F508$  mutations on both copies of *CFTR*. This is the only study that has directly measured a copy number variable locus in CF patients.

## Alpha Defensins as a Modifier Locus

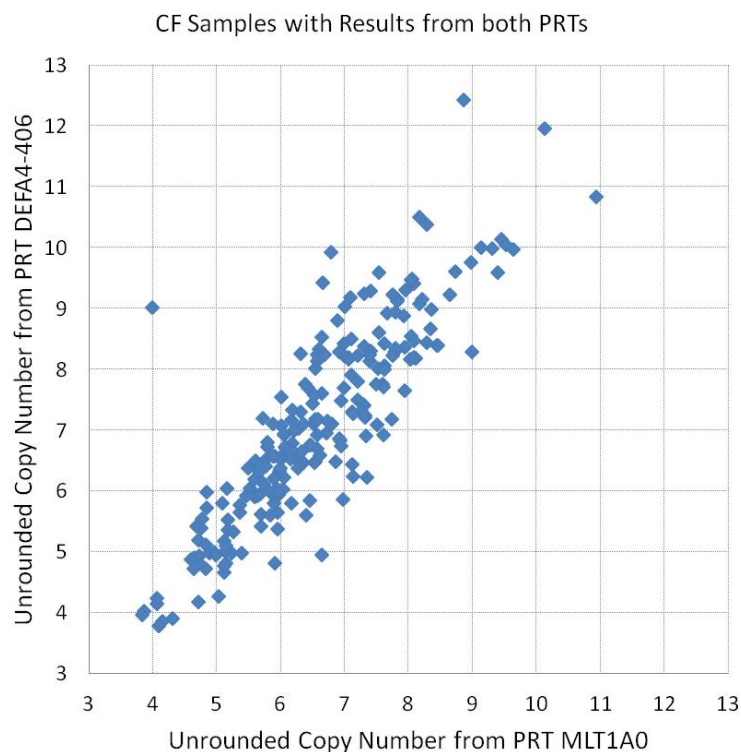
Neutrophil alpha defensins are not only antimicrobial but have been shown to be anti-inflammatory (Miles *et al.*, 2009). Given the infiltration of neutrophils in the CF lung, and the abundance of alpha defensins in neutrophils, both these properties could make them an important player in the lung pathogenesis of CF. Given the difficulty in studying CF lung disease development that has often led to contradictory results for the same questions, there is no obvious direction to a prior hypothesis that a low or a high copy number of *DEFA1A3* would enhance or decrease lung function, if at



all. In this study, the effect of *DEFA1A3* CNV on CF lung disease severity has been explored through two routes:

### A. *DEFA1A3* Copy Number Measurement in CF Patients

This is a direct study of the alpha defensin CNV in a cohort of 270 UK CF patients in collaboration with Dr. Jane Davies (Imperial College London). Lung function data has been collected for these patients allowing their grouping into mild and severe disease, and this information is available to Dr. Davies and her group only. Anonymized DNA samples were received for *DEFA1A3* analysis. Copy number measurements were done using PRT MLT1A0, PRT DEFA4-406, DefHae3 and Indel-5 assays. SNP rs4300027 was also genotyped. The samples used as calibrators for the PRTs were HapMap CEPH samples that were also part of the 3-generation CEPH families typed and there was high confidence in their measured *DEFA1A3* copy numbers. The graph below shows the level of concordance between the measured copy numbers from the two PRTs (231 samples had results from both PRTs).



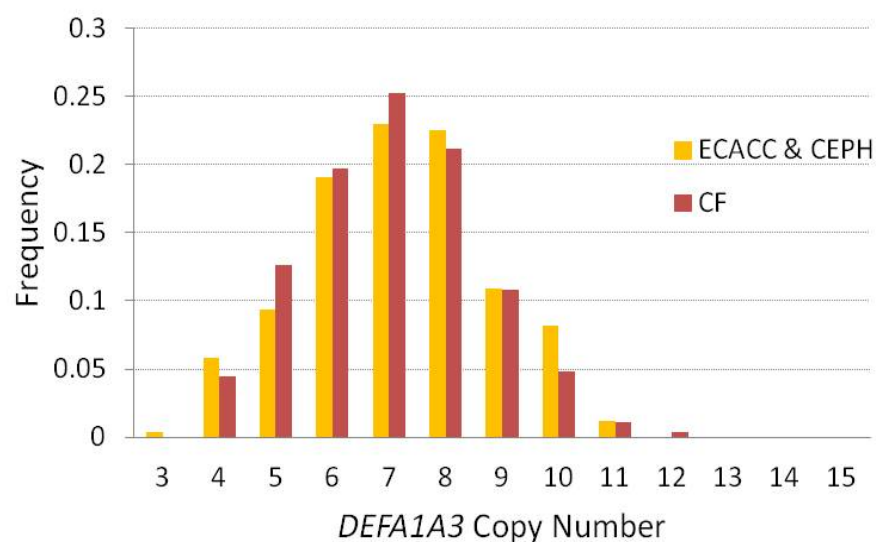
**Figure 6.10** Comparison of copy numbers measured from the two PRTs for 231 CF samples.

It was noticed that copy numbers from PRT DEFA4-406 tended to be higher than those from PRT MLT1A0 (Figure 6.10). In such cases, copy numbers

measured from PRT DEFA4-406 were more often closer to the MLCN (about 81% of times), which takes into account ratios from the Indel-5 and DefHae3 assays, than PRT MLT1A0 copy numbers, indicating a likelihood of bias in errors, with more from PRT MLT1A0. A correction to all PRT MLT1A0 measurements was tested, but it did not improve the overall accuracy of copy number measurements, as determined by the minimum ratios from the MLCN program (see Section 2.4) and copy number correlation with rs4300027 genotypes. Thus, the correction was not applied.

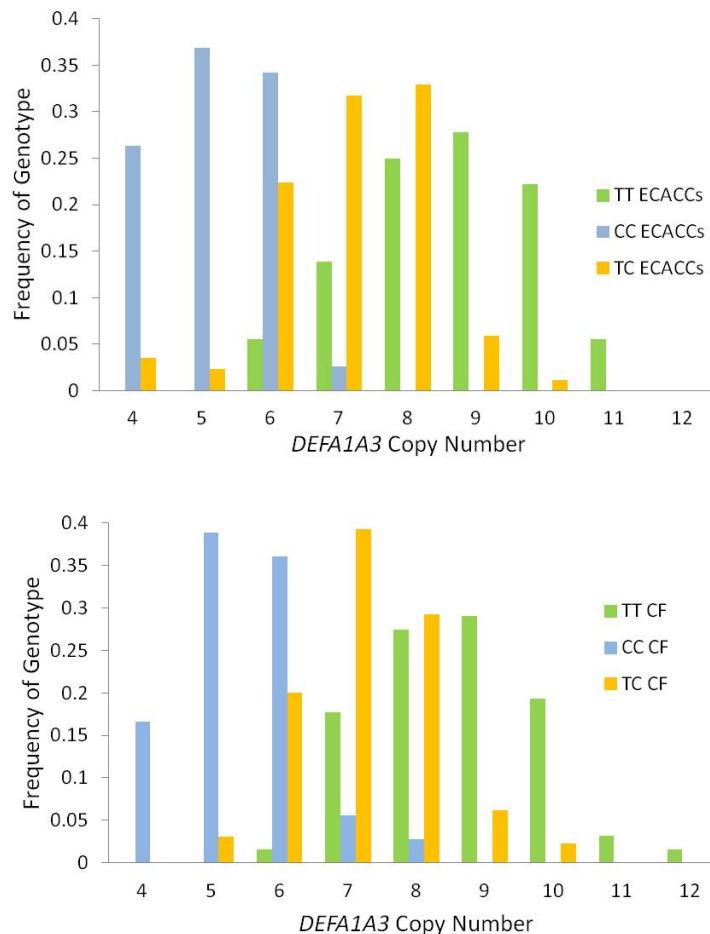
The copy number measured for one sample from PRT DEFA4-406 (9 copies) was about twice as high as that from PRT MLT1A0 (4 copies) (Figure 6.10). This was tested and found to be a carrier for the *DEFA4* gene deletion (see Section 2.10).

The overall copy number range and distribution for the 270 CF samples did not vary significantly (T-test p-value= 0.99) from the general population (ECACC and CEPH samples) as was expected (Figure 6.11).



**Figure 6.11** Comparison of the range and frequency of measured *DEFA1A3* copy numbers from CF patients and normal controls

SNP rs4300027 genotypes show the same correlation with copy number as had been observed in the general population (ECACCs):



**Figure 6.12** Comparison of rs4300027 genotype frequency distribution in *DEFA1A3* copy number classes of ECACC samples (top graph) and CF samples (bottom graph)

## Results of Analysis

Copy number data (total and for *DEFA3*) and rs4300027 genotypes for the 270 CF samples was sent to Dr. Davies for analysis with patient phenotype. They found no significant association of FEV<sub>1</sub> with total copy number (ANOVA p-value 0.17), with *DEFA3* copy number (ANOVA p-value 0.6), or with genotype (ANOVA p-value 0.7). They also found no significant association of genotype with pancreatic sufficiency (chi-squared p-value 0.96), with *Staphylococcus aureus* infection (chi-squared p-value 0.1), with MRSA infection (chi-squared p-value 0.36), with *Pseudomonas aeruginosa* infection (chi-squared p-value 0.97), with *Stenotrophomonas maltophilia* infection (chi-squared p-value 0.76), with *Aspergillus fumigatus* infection (chi-squared p-value 0.27) or with allergic bronchopulmonary aspergillosis (chi-squared p-value 0.9).

## B. SNPs as Copy Number Surrogates

The other approach to look for association between *DEFA1A3* copy number and CF lung disease severity was to exploit existing SNP genotypes from a

large-scale GWAS, by examining p-values obtained for copy number-tagging SNPs (4.3) (Wright *et al.*, 2011). Table 6.1 shows the p-values obtained.

**Table 6.1** List of SNPs and the haplotype features they potentially tag, and their p-values in the Cystic Fibrosis GWAS

SNP	Haplotypes Tagged	p-value for $\Delta F508$ homozygotes*	p-value for all samples**
rs4300027	2 and 3-copy versus 4 and 5-copy	0.05	0.1
rs4512398	as above	0.06	0.12
rs2702852	<i>DEFA3</i> presence/absence	0.02	0.02
rs2738058	as above	1	1
rs6984215	5-copy versus others	0.13	.09

\*1,978 individuals

\*\*includes 1,978  $\Delta F508$  homozygotes and 516 other CF genotypes

rs4300027 is the only SNP for which the LD with copy number haplotypes has been tested and replicated in ECACC samples (see Section 4.3). rs4512398 is in perfect or almost perfect LD with it, and so both SNPs result in similar p-values in this study. However, the p-values are not low enough (in the  $10^{-3}$  range) to suggest a significant effect of total *DEFA1A3* copy number on lung disease variability. *DEFA3* presence/absence tagging SNPs (rs2702852 and rs2738058) have been chosen from their apparent association observed in only 80 CEPH haplotypes (p-values  $3.9 \times 10^{-3}$  and  $7.8 \times 10^{-3}$ ) and rs6984215 from 56 CEPH haplotypes (p-value 0.013). There is no independent verification of their association and any could be a false positive. The lowest p-value observed is 0.02 for rs2702852 which has an apparent *DEFA3* gene presence tagging ability. The other SNP chosen for the same haplotype feature, rs2738058 is in weak LD with rs2702852 ( $R^2=0.5$  from 120 haplotypes). Thus either one or none of the two SNPs actually tag *DEFA3* gene-containing haplotypes, but not both. If rs2702852 does tag *DEFA3* gene presence/absence, and if it has a significant effect on CF lung disease severity, the low p-value (0.02) would suggest an effect that is so weak as to be of no further interest. The SNP to achieve genome-wide significance in this GWAS gave a p-value of  $3.3 \times 10^{-8}$  and accounted for 1 to 2% of variation in  $\Delta F508$  homozygous individuals.

## 6.7. Conclusion and Discussion

The protocol of neutrophil separation and subsequent protein extraction, adapted from Linzmeier and Ganz (Linzmeier *et al.*, 2005), has been shown to work well, followed by a crude separation of the defensin peptides on a small scale SDS PAGE gel that was used for sequencing them and a clear separation on a larger gel. Given their small size and basic nature, neutrophil defensins can be separated from other proteins and from each other (despite a single amino acid difference between DEFA1, 2 and 3) on an acid urea gel. This separation of the three defensins may even allow quantification of each separately via quantifying band intensity on the gel. The method of quantification used in this study is ELISA, which cannot differentiate between DEFA1, 2 and 3, and so total quantities are measured. However, this has been a preliminary study where only the protocol has been established and needs to be expanded to a larger scale for studying correlation between gene copy number of *DEFA1A3* and expression levels. It is crucial that the same number of cells from each tested individual be used for protein quantification. There are several steps where an error in cell count can be introduced, and while error is not completely unavoidable, reproducibility of results is necessary between samples. The first step is counting the separated neutrophils using a haemocytometer and subsequent aliquots of  $10^5$  cells are made based on this count. Counting cells manually can vary from person to person despite adopting the same technique, and hence should be made as uniform as possible. Second, when these aliquots are centrifuged to pellet the cells and remove the supernatant, the quality of cell pellet should be uniform between samples and also the removal of supernatant. After suspending the cells in acid, they are left overnight and then vacuum-dried to get dry protein extracts. These extracts are suspended in 1 mL of sample buffer and at this stage samples should be mixed uniformly and thoroughly, as it has been observed that portions of the dry pellet can tend to remain intact. For performing ELISA, it would be ideal if these suspensions are the correct dilution required for testing rather than having to further dilute them, which could potentially introduce further error. In the ELISA experiment (Figure 6.8 and Figure 6.9), the range of concentrations covered by the standards is from 0.07  $\mu\text{g/ml}$  to 10  $\mu\text{g/ml}$ . The previously measured concentration of neutrophil defensins is 1-5  $\mu\text{g}/10^6$  neutrophils. Thus, using extracts from  $10^5$  neutrophils suspended in 1 mL gave the correct dilution range to be used. However, note that the lowest value measured is about 0.13  $\mu\text{g/ml}$  for the 7-copy sample. If it is assumed that lower-copy samples, or indeed samples with any number of copies, could give lower expression levels, they would fall outside the measuring range of this experiment. Thus, either standard concentrations or

sample dilutions may have to be adjusted further to include all samples in the measurable range.

To investigate whether *DEFA1A3* copy number affects lung disease severity in Cystic Fibrosis, two approaches have been used. In one, CF samples have been typed by the four-assay system for copy number and this data has been compared with the patients' disease severity measures by our collaborators. Lung disease severity (measured as FEV<sub>1</sub>), pancreatic sufficiency and infection by several microorganisms, all gave no significant association with *DEFA1A3* copy number. In the other, as was done in investigating other GWAS, p-values for copy number haplotype-tagging SNPs have been requested from a CF GWAS. Given the power of this study, we can take the observation that none of the p-values is significant as a reliable indication that there is no effect of alpha-defensin copy number on CF severity strong enough to warrant further investigations.

## 7. DISCUSSION

### 7.1. Measurement of multiallelic CNVs

Chapters 3 and 4 detail the results from and the accuracy of the PRT-based measurement system. The use of reliable calibrators, agreement between the four assays and the observation of mendelian segregation in families are characteristics of this study that give confidence in the copy number measurements. Additionally, agreement with an external, independent assay (see Section 3.2) has also been demonstrated. The combination of these factors is not seen in any other study for *DEFA1A3* measurement. While this system has worked well in the European population, its global application is restricted either by population-specific SNPs under primers or differences in quality of DNA samples. The latter was observed in typing Cystic Fibrosis samples of European origin, which gave poor results when calibrated with the eight standard samples, but better-quality results when calibrated with HapMap CEPH samples. This could be because of either differences in how each DNA sample was prepared, or due to a general decline in the quality of DNA with time. The standard DNA samples used were older and had been through numerous freeze-thaw cycles, as compared to the HapMap CEPH samples.

When considering the general applicability of PRT for CNV measurement, it is greatly restricted by the availability of appropriate paralogous sequences. What other methodologies offer an advantage over these limitations? Locus-specific studies for *DEFA1A3* are limited to Aldred et al (Aldred *et al.*, 2005) and Linzmeier et al (Linzmeier *et al.*, 2005) which use MAPH and real-time PCR respectively, and have their own shortcomings as has been discussed previously (see Section 1.2). More recent studies for copy number genotyping are genome-wide, either genome-wide CNV arrays (array CGH) or whole-genome sequencing using next-generation technologies. Genome-wide CNV array study with the highest resolution (smallest probe size) is that of Conrad et al (Conrad *et al.*, 2010). As shown by the comparison of their results with the PRT-based copy numbers from this study (see Section 3.2), and especially the tight agreement between the expected and obtained  $\log_2$  hybridization ratios, the accuracy of their results for *DEFA1A3* is no less than the PRT-based system. However, they could not assign *DEFA1A3* copy numbers to the tested samples because they did not know the copy number for the reference sample or have any calibrating samples. This study, being genome-wide, was

meant to characterize as many CNVs as it could, and highly variable CNVs like *DEFA1A3*, if they remained intractable to their approach, were not pursued further. This is true for all genome-wide approaches. Their concern is obtaining a global picture rather than pursuing individual loci.

So how useful or reliable can these emerging technologies be for genotyping multiallelic CNVs? A next-generation sequencing study for measuring CNVs using read depth by Sudmant et al (Sudmant *et al.*, 2010) has been briefly discussed previously (Section 1.2). Unlike the array-based studies where the relative hybridization signal from two samples is used to infer copy number, which makes it necessary to know the copy number of one of those samples at any given locus, genome-wide sequencing makes use of read depth, relying on the fact that largely the genome is only in two copies and any variation from this can be measured and assigned copy numbers. From the whole genome there will be instances of all copy number classes, 0 to hundreds, and if read depth is an accurate measure, they can be assigned without any prior knowledge of copy numbers in any sample. However, the amplification of the short reads prior to sequencing can introduce a bias, such that the representation of the reads may not accurately mirror their relative abundance in the test genome. This may systematically affect some loci more than others and could also be stochastic, such that repeated sequencing of the same genome on different occasions may lead to different results for the same loci (batch effect) (Taub *et al.*, 2010). This batch effect is mainly due to sampling variability. If we consider sampling to be a random process it should follow the Poisson distribution. A 10x coverage of a genome in a sequencing experiment would mean that a 7-copy repeat sequence, 10kb in size, should on average have 7000 reads ideally, if the average read length is 100 bp. With such high reads, the probability of assigning a copy number other than 7 is negligibly low. However, for copy-variable regions, such small read lengths mean that actual sequences that can be unambiguously assigned to their position in the genome are much fewer than the entire variable sequence. In the case of the alpha defensin locus, for example, there will be ambiguity between the full and partial repeats, between the *DEFA1/3* genes and *DEFA4* gene and between other surrounding paralogous sequences or dispersed repeats. This reduces coverage by a magnitude depending upon the CNV locus in question. To overcome these inconsistencies, an approach using 'unique molecular identifiers' has been developed (Kivioja *et al.*, 2011). The attachment of synthetic DNA barcodes to a fragmented sample prior to its amplification allows the counting of sequences via observation of unique barcodes rather than read depth. It has been shown to work well for whole chromosome copy numbers (aneuploidies, X and Y chromosomes etc.) with a



higher accuracy than the read depth method. For whole chromosome copy numbers, DNA equivalent to one genome was sampled as a starting material. This dilution approach is probably not feasible for CNVs in the size range of kilobase pairs due to sampling bias, unless the same genomic sample is tested several times.

It remains to be seen what combination of approaches would allow accurate CNV measurement for submicroscopic variant loci, such as *DEFA1A3*. More generally, it would probably be a combination of several approaches (genome-wide and locus specific), depending upon the locus and its characteristics, which would allow large-scale and accurate CNV genotyping of the level that is required for further structural and functional studies.

## **7.2. Evolution of *DEFA1A3* Haplotypes**

Using haplotype information for *DEFA1A3* copy numbers inferred from segregation in CEPH families and combining it with genotype data of surrounding SNPs has allowed a view of haplotypic structure in Europeans. SNPs have been found that tag copy number lineages and also other features of the locus, e.g. presence of the *DEFA3* gene or the Indel-5 minor allele. However, only one SNP association with copy number haplotypes, that of rs4300027, has been validated, and it is not found in non-European populations. In the Japanese population, although only diploid copy numbers were available for analysis, it nevertheless has shown the presence of distinct haplotypes that allow copy number tagging by SNPs. The Chinese and African populations do not show any strong LD between SNPs and copy number classes of repeats. Without further knowledge on the locus, several plausible evolutionary paths can be constructed to explain these observations. Simultaneously, characteristics of this locus can be inferred, for example the rate of copy number change.

In the absence of selection, limited variation in the Europeans and Japanese shows the limit of the copy number mutation rate. For a multiallelic locus such as *DEFA1A3*, recombination between haplotypes would be considered the main mechanism giving rise to copy number changes. If recombination changed copy numbers often enough then LD with surrounding markers would have broken down. Of course recombination could also occur without changing copy numbers, but still affecting linkage with surrounding markers. It could also result in gene conversion events within the repeats, again without affecting copy numbers. As stated in Chapter 4, not observing any copy number change in the number of transmissions in the CEPH families studied allows a tentative upper limit for copy number-changing rate to be estimated

( $8.86 \times 10^{-3}$  mutations per meiosis, 95% CI). It cannot be said how low it actually might be.

This restricted variation in the Europeans and Japanese may have come about due to a bottleneck effect and further genetic drift. When considering this locus in terms of number of repeats alone, assuming no selection, and looking at the African and Chinese populations only, the observed frequency of copy number classes can be explained by a balance between mutations that increase copy number and those that decrease it. This idea would be similar to the phenomenon of equilibrium allele frequency distribution in microsatellites, although the mechanism of copy number change would not be the same. Given the strong bottleneck effect observed in the Europeans and Japanese, as evidenced by their LD structure, they are excluded from this consideration.

In the discussion so far, the possibility of selective forces acting upon this locus has been ignored. A lot can only be hypothesized regarding selection because it is unknown how important variation in defensin levels can be for human health or even if expression varies significantly according to the haplotypes a person possesses. For simplicity we assume that defensin levels are affected by genomic variation. We must consider not only variation in terms of number of repeats, but also in terms of sequence variants within those repeats that might affect expression. For example, this would mean that not all 5-copy haplotypes would give the same level of expression, but varying levels depending upon the 'type' of repeats making up the array. One scenario that would explain the current common European haplotypes could be that they were selected for by agents encountered by early Europeans on the basis of the number of *functional* repeats or variants that affect expression levels. For example, this would explain how all 5-copy haplotypes observed in CEPHS have allele T at SNP rs4300027, whereas non-Europeans have both C and T. The strength of the selective pressure would dictate how readily some alleles would be lost and others become frequent. The same kind of hypothesis could be applied to the Japanese, where high-copy haplotypes have become much more common, and can be taken as an indication for higher defensin levels to be beneficial in the environments faced by that population. Selection could also explain the general copy number frequency distribution such that a median number of defensins (6, 7 or 8) is beneficial, while lower numbers are harmful/less beneficial under some selective forces and higher numbers are harmful under others. However, this would not be true for the Japanese population where the copy number frequency distribution does not follow the normal-like distribution seen in other populations. Another possibility in the European and/or Japanese populations is that selection might have occurred

on another locus in the same LD region, and that is the *DEFA5* gene. This would indirectly limit variation or make certain haplotypes more common.

Another important aspect of the European haplotypes that remains unexplored in this study is the arrangement of repeats along the array. Is the *DEFA3* gene always present in the partial repeat? Is the Indel-5 minor allele present on repeats with a *DEFA1* gene? Answers to these questions will build a finer picture of haplotypic diversity, which in turn will help to understand the evolution of the locus. For example, it has been observed that 5-copy haplotypes largely carry only a single copy of the Indel-5 minor allele (28 out of 34). So in all those haplotypes is it always the same repeat that carries the Indel-5 minor allele or is it several different positions of repeats? If it is the first instance, it builds a picture of lower variation than if it is the latter case. We already know of a variant feature that splits these haplotypes into two groups, and that is that six out of the twenty-eight have no *DEFA3* while the rest have at least one copy of it. However, it does not mean that these haplotypes must differ in terms of the Indel-5 minor allele position. Additionally, even if they all did carry the same arrangement of repeats with respect to the Indel-5 variant, they could still be a case of identity-by-state rather than identity-by-descent. However, since these features seem to be well-tagged by SNPs in the haplotypes studied, they are more likely to be identical-by-descent.

Whatever the mechanism was that led to the restricted variation in the Europeans, it predicts that further definition of the haplotypes in terms of types of repeats depending upon sequence variants within them and their position in the arrays should again show a limited number of haplotypes that are tagged by surrounding SNPs. Ultimately, it might be possible to define a hierarchical set of SNPs, probably beginning with rs4300027, whose genotypes will allow us to follow much more well-defined haplotypes than just total copy number haplotypes.

### **7.3. *DEFA1A3* CNV and Clinical Phenotypes**

In this study, the effect of copy number variation of the alpha defensin genes has been investigated in several immunity-related diseases via genotypes of SNPs that tag copy number lineages, and no significant association has been found. There was no prior indication for these defensins to affect any of the phenotypes studied, or any not investigated. How important a role do defensins play normally? Neutrophils store huge amounts of these peptides, but also have other antimicrobial peptides and oxidative killing mechanisms.

In two known conditions of neutrophil defensin deficiency, Specific Granule Deficiency (SGD) and Chediak-Higashi Syndrome (CHS), it is not just the defensins that are deficient but also other neutrophil granule components, like elastase, myeloperoxidase and lactoferrin, all with antimicrobial functions. In both diseases, patients suffer from immune deficiency and recurrent infections. Another disease of phagocyte insufficiency is Chronic Granulomatous Disease (CGD) which is caused by reduced or no production of superoxide radicals in the neutrophils, monocytes and macrophages. CGD patients are highly susceptible to infections from bacteria and fungi. This indicates that the absence of antimicrobial mechanisms in the neutrophil other than the defensins is sufficient to cause immune deficiency, which diminishes the importance of defensins despite their large quantities. It would be interesting to see if defensin gene copy number would affect disease severity in CGD, but since it is a rare and heterogenous group of diseases (incidence estimated at 1 in 250,000), large sized samples matched for genetic mutation and available clinical data will be difficult to obtain.

The effect of gene copy number on protein expression also remains to be established. If say expression was proportional to the number of functional *DEFA1A3* copies, then it would not be the total copy number, but the copy number of functional repeats that would matter. Given that this locus is highly variable in copy number in the normal population, it would not be surprising to find haplotypes where only some or even none of the repeats produces a functional copy of the peptides. Thus, when correlating gene copy numbers obtained from PRT-based measurements with protein expression levels, if the results reveal variation in protein levels but not in a coherent manner relative to copy number, it might be indicative of sequence variants affecting expression. This situation would also render rs4300027 useless as a proxy for copy number in functional studies.

Spatial arrangement of the repeats may also be important if expression depended upon position. In this possible scenario, expression may be restricted to a particular repeat position, e.g. the partial repeat at the centromeric end, or the full repeat at the telomeric end would be the only one to be expressed. This would give each individual two transcribed copies which may have varying levels of expression based upon sequence variants. However, previous studies have shown not only that protein levels vary between people (Linzmeier *et al.*, 2005), but also that mRNA levels of *DEFA1* and *DEFA3* have the same ratio as does the *DEFA1* and *DEFA3* gene numbers (Aldred *et al.*, 2005). This indicates that at least most, if not all repeats are transcriptionally active. Aldred *et al.* also found that the level of mRNA expression from *DEFA3* genes was twice that from *DEFA1* genes. It will be

interesting to see if this replicates in terms of protein levels. ELISA will not allow separate quantification of the three peptides, but gel band intensity on an acid urea gel might. However, either or both of DEFA1 and DEFA3 peptides are converted to DEFA2 which could alter their ratios.

Results from studying the correlation between gene numbers and protein levels and defining the haplotypes at a finer level of sequence variance and spatial arrangement will be informative for each other. While variation in the repeat haplotypes might help explain variation in expression alongside unravelling evolutionary history, how expression correlates with copy number will help in understanding the functional significance and potential selective forces (or absence of) that might have affected the evolution of this locus.

## 8. Bibliography

(2001). "Initial sequencing and analysis of the human genome." Nature **6822**(409): 860-921

S. Abu Bakar, E. J. Hollox and J. A. L. Armour (2009). "Allelic recombination between distinct genomic locations generates copy number diversity in human beta-defensins." Proceedings of the National Academy of Sciences of the United States of America **3**(106): 853-858

A. Alaedini and P. H. R. Green (2005). "Narrative review: Celiac disease: Understanding a complex autoimmune disorder." Annals of Internal Medicine **4**(142): 289-298

N. Aldallal, E. E. McNaughton, L. J. Manzel, A. M. Richards, J. Zabner, et al. (2002). "Inflammatory response in airway epithelial cells isolated from patients with cystic fibrosis." American Journal of Respiratory and Critical Care Medicine **9**(166): 1248-1256

M. C. Aldhous, S. Abu Bakar, N. J. Prescott, R. Palla, K. Soo, et al. (2010). "Measurement methods and accuracy in copy number variation: failure to replicate associations of beta-defensin copy number with Crohn's disease." Human Molecular Genetics **24**(19): 4930-4938

P. M. R. Aldred, E. J. Hollox and J. A. L. Armour (2005). "Copy number polymorphism and expression level variation of the human alpha-defensin genes DEFA1 and DEFA3." Human Molecular Genetics **14**(14): 2045-2052

D. M. Altshuler, R. A. Gibbs, L. Peltonen, E. Dermitzakis, S. F. Schaffner, et al. (2010). "Integrating common and rare genetic variation in diverse human populations." Nature **7311**(467): 52-58

P. D. Arkwright, S. Laurie, M. Super, V. Pravica, M. J. Schwarz, et al. (2000). "TGF-beta(1) genotype and accelerated decline in lung function of patients with cystic fibrosis." Thorax **6**(55): 459-462

J. A. Armour, C. Sismani, P. C. Patsalis and G. Cross (2000). "Measurement of locus copy number by hybridisation with amplifiable probes." Nucleic Acids Res **2**(28): 605-609.

J. A. L. Armour, R. Palla, P. L. J. M. Zeeuwen, M. den Heijer, J. Schalkwijk, et al. (2007). "Accurate, high-throughput typing of copy number variation using paralogue ratios from dispersed repeats." Nucl. Acids Res. **3**(35): e19

- E. Ballana, J. R. Gonzalez, N. Bosch and X. Estivill (2007). "Inter-population variability of DEFA3 gene absence: Correlation with haplotype structure and population variability." BMC Genomics 8):
- R. Bals and P. S. Hiemstra (2004). "Innate immunity in the lung: how epithelial cells fight against respiratory pathogens." European Respiratory Journal 2(23): 327-333
- J. C. Barrett, S. Hansoul, D. L. Nicolae, J. H. Cho, R. H. Duerr, et al. (2008). "Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease." Nature Genetics 8(40): 955-962
- A. D. Befus, C. Mowat, M. Gilchrist, J. Hu, S. Solomon, et al. (1999). "Neutrophil defensins induce histamine secretion from mast cells: Mechanisms of action." Journal of Immunology 2(163): 947-953
- R. W. Bentley, J. Pearson, R. B. Gearry, M. L. Barclay, C. McKinney, et al. (2010). "Association of Higher DEFB4 Genomic Copy Number With Crohn's Disease." American Journal of Gastroenterology 2(105): 354-359
- P. Birrer, N. G. McElvaney, A. Rudeberg, C. W. Sommer, S. Liechtigallati, et al. (1994). "Protease-Antiprotease Imbalance in the Lungs of Children with Cystic Fibrosis." American Journal of Respiratory and Critical Care Medicine 1(150): 207-213
- R. C. Boucher (2004). "New concepts of the pathogenesis of cystic fibrosis lung disease." European Respiratory Journal 1(23): 146-158
- A. M. Bowcock and J. G. Krueger (2005). "Getting under the skin: the immunogenetics of psoriasis." Nat Rev Immunol 9(5): 699-711
- K. A. Brogden (2005). "Antimicrobial peptides: Pore formers or metabolic inhibitors in bacteria?" Nature Reviews Microbiology 3(3): 238-250
- C. E. G. Bruder, A. Piotrowski, A. Gijssbers, R. Andersson, S. Erickson, et al. (2008). "Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles." American Journal of Human Genetics 3(82): 763-771
- R. E. Brugha and J. C. Davies (2011). "Pseudomonas aeruginosa in cystic fibrosis: pathogenesis and new treatments." British Journal of Hospital Medicine 11(72): 614-619
- F. Capon, M. J. Bijlmakers, M. Quaranta, K. Timms, A. D. Burden, et al. (2008). "Characterization of ZNF313/RNF114 as a novel psoriasis susceptibility gene." British Journal of Dermatology 6(159): 1389-1389

- O. Chertov, D. F. Michiel, L. L. Xu, J. M. Wang, K. Tani, et al. (1996). "Identification of defensin-1, defensin-2, and CAP37/azurocidin as T-cell chemoattractant proteins released from interleukin-8-stimulated neutrophils." Journal of Biological Chemistry **6**(271): 2935-2940
- J. H. Cho (2008). "The genetics and immunopathogenesis of inflammatory bowel disease." Nature Reviews Immunology **6**(8): 458-466
- Y. Colin, B. Cherifzahar, C. L. Vankim, V. Raynal, V. Vanhuffel, et al. (1991). "Genetic-Basis of the Rhd-Positive and Rhd-Negative Blood-Group Polymorphism as Determined by Southern Analysis." Blood **10**(78): 2747-2752
- D. F. Conrad, D. Pinto, R. Redon, L. Feuk, O. Gokcumen, et al. (2010). "Origins and functional impact of copy number variation in the human genome." Nature **7289**(464): 704-712
- N. Craddock, M. E. Hurles, N. Cardin, R. D. Pearson, V. Plagnol, et al. (2010). "Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls." Nature **7289**(464): 713-U786
- A. Craig, J. Mai, S. S. Cai and S. Jeyaseelan (2009). "Neutrophil Recruitment to the Lungs during Bacterial Pneumonia." Infection and Immunity **2**(77): 568-575
- R. N. Cunliffe (2003). "alpha-Defensins in the gastrointestinal tract." Molecular Immunology **7**(40): 463-467
- K. A. Daher, M. E. Selsted and R. I. Lehrer (1986). "Direct Inactivation of Viruses by Human Granulocyte Defensins." Journal of Virology **3**(60): 1068-1074
- K. A. Daher, R. I. Lehrer, T. Ganz and M. Kronenberg (1988). "Isolation and characterization of human defensin cDNA clones." Proceedings of the National Academy of Sciences of the United States of America **19**(85): 7327-7331
- M. J. Daly, J. D. Rioux, S. E. Schaffner, T. J. Hudson and E. S. Lander (2001). "High-resolution haplotype structure in the human genome." Nature Genetics **2**(29): 229-232
- S. Das, N. Nikolaidis, H. Goto, C. McCallister, J. X. Li, et al. (2010). "Comparative Genomics and Evolution of the Alpha-Defensin Multigene Family in Primates." Molecular Biology and Evolution **10**(27): 2333-2343
- Y. Date, M. Nakazato, K. Shiomi, H. Toshimori, K. Kangawa, et al. (1994). "Localization of Human Neutrophil Peptide (HNP) and its Messenger RNA in Neutrophil Series." Annals of Hematology **2**(69): 73-77



- P. B. Davis (2006). "Cystic fibrosis since 1938." American Journal of Respiratory and Critical Care Medicine **5**(173): 475-482
- A. De Benedetto, R. Agnihothri, L. Y. McGirt, L. G. Bankova and L. A. Beck (2009). "Atopic Dermatitis: A Disease Caused by Innate Immune Defects?" Journal of Investigative Dermatology **1**(129): 14-30
- P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, et al. (2001). "Initial sequencing and analysis of the human genome (vol 409, pg 860, 2001)." Nature **6846**(412): 565-566
- B. B. A. de Vries, R. Pfundt, M. Leisink, D. A. Koolen, L. Vissers, et al. (2005). "Diagnostic genome profiling in mental retardation." American Journal of Human Genetics **4**(77): 606-616
- G. Diamond, N. Beckloff, A. Weinberg and K. O. Kisich (2009). "The Roles of Antimicrobial Peptides in Innate Host Defense." Current Pharmaceutical Design **21**(15): 2377-2392
- B. Diosdado, H. Van Bakel, E. Strengman, L. Franke, E. Van Oort, et al. (2007). "Neutrophil recruitment and barrier impairment in celiac disease: A genomic study." Clinical Gastroenterology and Hepatology **5**(5): 574-581
- S. J. Diskin, C. P. Hou, J. T. Glessner, E. F. Attiyeh, M. Laudenslager, et al. (2009). "Copy number variation at 1q21.1 associated with neuroblastoma." Nature **7249**(459): 987-U112
- C. M. Doerschuk, N. Beyers, H. O. Coxson, B. Wiggs and J. C. Hogg (1993). "Comparison of Neutrophil and Capillary Diameters and their Relation to Neutrophil Sequestration in Lungs." Journal of Applied Physiology **6**(74): 3040-3045
- G. Doring, H. Kroghjohansen, S. Weidinger and N. Hoiby (1994). "Allotypes of Alpha(1)-Antitrypsin in Patients with Cystic Fibrosis, Homozygous and Heterozygous for DeltaF508." Pediatric Pulmonology **1**(18): 3-7
- V. Driss, F. Legrand, E. Hermann, S. Loiseau, Y. Guerardel, et al. (2009). "TLR2-dependent eosinophil interactions with mycobacteria: role of alpha-defensins." Blood **14**(113): 3235-3244
- N. Droin, A. Jacquiel, J. B. Hendra, C. Racoeur, C. Truntzer, et al. (2010). "Alpha-defensins secreted by dysplastic granulocytes inhibit the differentiation of monocytes in chronic myelomonocytic leukemia." Blood **1**(115): 78-88

- M. L. Drumm, M. W. Konstan, M. D. Schluchter, A. Handler, R. Pace, et al. (2005). "Genetic modifiers of lung disease in cystic fibrosis." New England Journal of Medicine **14**(353): 1443-1453
- P. C. A. Dubois, G. Trynka, L. Franke, K. A. Hunt, J. Romanos, et al. (2010). "Multiple common variants for celiac disease influencing immune gene expression." Nature Genetics **4**(42): 295-U242
- B. Ericksen, Z. B. Wu, W. Y. Lu and R. I. Lehrer (2005). "Antibacterial activity and specificity of the six human alpha-defensins." Antimicrobial Agents and Chemotherapy **1**(49): 269-275
- K. Fellermann, D. E. Stange, E. Schaeffeler, H. Schmalzl, J. Wehkamp, et al. (2006). "A Chromosome 8 Gene-Cluster Polymorphism with Low Human Beta-Defensin 2 Gene Copy Number Predisposes to Crohn Disease of the Colon." Am. J. Hum. Genet. **79**: 439-448
- S. F. Field, J. M. M. Howson, L. M. Maier, S. Walker, N. M. Walker, et al. (2009). "Experimental aspects of copy number variant assays at CCL3L1." Nature Medicine **10**(15): 1115-1117
- P. Fode, C. Jespersgaard, R. J. Hardwick, H. Bogle, M. Theisen, et al. (2011). "Determination of Beta-Defensin Genomic Copy Number in Different Populations: A Comparison of Three Methods." Plos One **2**(6):
- D. D. Frangolias, J. Ruan, P. J. Wilcox, G. F. Davidson, L. T. K. Wong, et al. (2003). "alpha(1)-antitrypsin deficiency alleles in cystic fibrosis lung disease." American Journal of Respiratory Cell and Molecular Biology **3**(29): 390-396
- T. Ganz, M. E. Selsted, D. Szklarek, S. S. L. Harwig, K. Daher, et al. (1985). "Defensins - Natural peptide antibiotics of human neutrophils." Journal of Clinical Investigation **4**(76): 1427-1435
- T. Ganz (1987). "Extracellular release of antimicrobial defensins by human polymorphonuclear leukocytes." Infection and Immunity **3**(55): 568-571
- T. Ganz (2003). "Defensins: antimicrobial peptides of innate immunity." Nat Rev Immunol **9**(3): 710-720
- I. Garcia-Verdugo, D. Descamps, M. Chignard, L. Touqui and J. M. Sallenave (2010). "Lung protease/anti-protease network and modulation of mucus production and surfactant activity." Biochimie **11**(92): 1608-1617
- D. Ghosh, E. Porter, B. Shen, S. K. Lee, D. Wilk, et al. (2002). "Paneth cell trypsin is the processing enzyme for human defensin-5." Nat Immunol **6**(3): 583-590

- R. A. Gibbs, J. W. Belmont, P. Hardenbol, T. D. Willis, F. L. Yu, et al. (2003). "The International HapMap Project." Nature **6968**(426): 789-796
- T. Gieseemann, G. Guttenberg and K. Aktories (2008). "Human alpha-defensins inhibit Clostridium difficile toxin B." Gastroenterology **7**(134): 2049-2058
- J. T. Glessner, K. Wang, G. Q. Cai, O. Korvatska, C. E. Kim, et al. (2009). "Autism genome-wide copy number variation reveals ubiquitin and neuronal genes." Nature **7246**(459): 569-573
- M. J. Goldman, G. M. Anderson, E. D. Stolzenberg, U. P. Kari, M. Zasloff, et al. (1997). "Human beta-defensin-1 is a salt-sensitive antibiotic in lung that is inactivated in cystic fibrosis." Cell **4**(88): 553-560
- M. J. Goldman, G. M. Anderson, E. D. Stolzenberg, U. P. Kari, M. Zasloff, et al. (1997). "Human beta-defensin-1 is a salt-sensitive antibiotic in lung that is inactivated in cystic fibrosis." Cell **4**(88): 553-560
- A. F. Gombart, M. Shiohara, S. H. Kwok, K. Agematsu, A. Komiyama, et al. (2001). "Neutrophil-specific granule deficiency: homozygous recessive inheritance of a frameshift mutation in the gene encoding transcription factor CCAAT/enhancer binding protein-epsilon." Blood **9**(97): 2561-2567
- J. E. Gordillo, S. Weidinger, R. R. Foelster-Holst, A. Bauerfeind, F. Ruschendorf, et al. (2009). "A Common Variant on Chromosome 11q13 is Associated with Atopic Dermatitis." Genetic Epidemiology **8**(33): 797-797
- J. P. Greer, Foerster, J., Rodgers, G. M., Paraskevas, F., Glader, B., Arber, D. A., Means Jr, R. T. (2009). *Wintrobe's Clinical Hematology*. (12). Philadelphia: Lippincott Williams & Wilkins.
- U. Griesenbach, D. M. Geddes and E. Alton (1999). "The pathogenic consequences of a single mutated CFTR gene." Thorax **54**: S19-S23
- T. W. L. Groeneveld, T. H. Ramwadhoebe, L. A. Trouw, D. L. van den Ham, V. van der Borden, et al. (2007). "Human neutrophil peptide-1 inhibits both the classical and the lectin pathway of complement activation." Molecular Immunology **14**(44): 3608-3614
- U. Gullberg, N. Bengtsson, E. Bulow, D. Garwicz, A. Lindmark, et al. (1999). "Processing and targeting of granule proteins in human neutrophils." Journal of Immunological Methods **1-2**(232): 201-210
- D. A. Hafler, A. Compston, S. Sawcer, E. S. Lander, M. J. Daly, et al. (2007). "Risk alleles for multiple sclerosis identified by a genomewide study." The New England journal of medicine **9**(357): 851-862

- H. Hakonarson, S. F. A. Grant, J. P. Bradfield, L. Marchand, C. E. Kim, et al. (2007). "A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene." Nature **7153**(448): 591-U597
- A. Hamosh and M. Corey (1993). "Correlation between Genotype and Phenotype in Patients with Cystic Fibrosis." New England Journal of Medicine **18**(329): 1308-1313
- J. Harder, J. Bartels, E. Christophers and J. M. Schroder (1997). "A peptide antibiotic from human skin." Nature **6636**(387): 861
- S. S. L. Harwig, A. S. K. Park and R. I. Lehrer (1992). "Characterization of Defensin Precursors in Mature Human Neutrophils." Blood **6**(79): 1532-1537
- P. J. Hastings, G. Ira and J. R. Lupski (2009). "A Microhomology-Mediated Break-Induced Replication Model for the Origin of Human Copy Number Variation." Plos Genetics **1**(5):
- C. N. Henrichsen, N. Vinckenbosch, S. Zollner, E. Chaignat, S. Pradervand, et al. (2009). "Segmental copy number variation shapes tissue transcriptomes." Nature Genetics **4**(41): 424-429
- D. A. Hinds, A. P. Kloek, M. Jen, X. Y. Chen and K. A. Frazer (2006). "Common deletions and SNPs are in linkage disequilibrium in the human genome." Nature Genetics **1**(38): 82-85
- E. J. Hollox, J. Davies, U. Griesenbach, J. Burgess, E. W. Alton, et al. (2005). "Beta-defensin genomic copy number is not a modifier locus for cystic fibrosis." J Negat Results Biomed **4**: 9
- E. J. Hollox, J. Davies, U. Griesenbach, J. Burgess, E. W. F. W. Alton, et al. (2005). "Beta-defensin genomic copy number is not a modifier locus for cystic fibrosis." J. Negative Results in BioMed. **4**: 9
- E. J. Hollox, U. Huffmeier, P. Zeeuwen, R. Palla, J. Lascorz, et al. (2008). "Psoriasis is associated with increased beta-defensin genomic copy number." Nature Genetics **1**(40): 23-25
- N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, et al. (2008). "Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays." Plos Genetics **8**(4):
- E. Houtmeyers, R. Gosselink, G. Gayan-Ramirez and M. Decramer (1999). "Regulation of mucociliary clearance in health and disease." European Respiratory Journal **5**(13): 1177-1188

- A. L. Hughes (1999). "Evolutionary diversification of the mammalian defensins." Cell Mol Life Sci **1-2**(56): 94-103.
- A. J. Iafrate, L. Feuk, M. N. Rivera, M. L. Listewnik, P. K. Donahoe, et al. (2004). "Detection of large-scale variation in the human genome." Nat Genet **9**(36): 949-951
- A. Itsara, G. M. Cooper, C. Baker, S. Girirajan, J. Li, et al. (2009). "Population Analysis of Large Copy Number Variants and Hotspots of Human Genetic Disease." American Journal of Human Genetics **2**(84): 148-161
- A. Itsara, H. Wu, J. D. Smith, D. A. Nickerson, I. Romieu, et al. (2010). "De novo rates and selection of large copy number variation." Genome Research **11**(20): 1469-1481
- T. Z. Khan, J. S. Wagener, T. Bost, J. Martinez, F. J. Accurso, et al. (1995). "Early Pulmonary Inflammation in Infants with Cystic Fibrosis." American Journal of Respiratory and Critical Care Medicine **4**(151): 1075-1082
- C. Kim, N. Gajendran, H. W. Mittrucker, M. Weiwad, Y. H. Song, et al. (2005). "Human alpha-defensins neutralize anthrax lethal toxin and protect against its fatal consequences." Proceedings of the National Academy of Sciences of the United States of America **13**(102): 4830-4835
- T. Kivioja, A. Vaharautio, K. Karlsson, M. Bonke, M. Enge, et al. (2011). "Counting absolute numbers of molecules using unique molecular identifiers." Nature Methods **1**(9): 72-U183
- R. J. Klein, C. Zeiss, E. Y. Chew, J. Y. Tsai, R. S. Sackler, et al. (2005). "Complement factor H polymorphism in age-related macular degeneration." Science **5720**(308): 385-389
- M. E. Klotman and T. L. Chang (2006). "Defensins in innate antiviral immunity." Nature Reviews Immunology **6**(6): 447-456
- M. R. Knowles, J. M. Robinson, R. E. Wood, C. A. Pue, W. M. Mentz, et al. (1997). "Ion composition of airway surface liquid of patients with cystic fibrosis as compared with normal and disease-control subjects." Journal of Clinical Investigation **10**(100): 2588-2595
- J. K. Kolls, P. B. McCray and Y. R. Chan (2008). "Cytokine-mediated regulation of antimicrobial proteins." Nature Reviews Immunology **11**(8): 829-835
- S. Kugathasan, R. N. Baldassano, J. P. Bradfield, P. M. A. Sleiman, M. Imielinski, et al. (2008). "Loci on 20q13 and 21q22 are associated with pediatric-onset inflammatory bowel disease." Nature Genetics **10**(40): 1211-1215

O. Lao, A. M. Andres, E. Mateu, J. Bertranpetit and F. Calafell (2003). "Spatial patterns of cystic fibrosis mutation spectra in European populations." European Journal of Human Genetics **5**(11): 385-394

V. LeCabec, J. B. Cowland, J. Calafat and N. Borregaard (1996). "Targeting of proteins to granule subsets is determined by timing and not by sorting: The specific granule protein NGAL is localized to azurophil granules when expressed in HL-60 cells." Proceedings of the National Academy of Sciences of the United States of America **13**(93): 6454-6457

A. Lehuen, J. Diana, P. Zacccone and A. Cooke (2010). "Immune cell crosstalk in type 1 diabetes." Nature Reviews Immunology **7**(10): 501-513

J. Lekstrom-Himes and K. G. Xanthopoulos (1998). "Biological role of the CCAAT enhancer-binding protein family of transcription factors." Journal of Biological Chemistry **44**(273): 28545-28548

J. A. Lekstrom-Himes, S. E. Dorman, P. Kopar, S. M. Holland and J. I. Gallin (1999). "Neutrophil-specific granule deficiency results from a novel mutation with loss of function of the transcription factor CCAAT enhancer binding protein epsilon." Journal of Experimental Medicine **11**(189): 1847-1852

R. M. Linzmeier and T. Ganz (2005). "Human defensin gene copy number polymorphisms: comprehensive analysis of independent variation in alpha- and beta-defensin regions at 8p22-p23." Genomics **4**(86): 423-430

Y. Liu, C. Helms, W. Liao, L. C. Zaba, S. Duan, et al. (2008). "A genome-wide association study of psoriasis and psoriatic arthritis identifies new disease loci." Plos Genetics **4**(4):

D. P. Locke, A. J. Sharp, S. A. McCarroll, S. D. McGrath, T. L. Newman, et al. (2006). "Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome." Am J Hum Genet **2**(79): 275 - 290

M. N. Madison, Y. Kleshchenko, P. Nde, K. Simmons, M. F. Lima, et al. (2007). "Human defensin alpha-1 kills Trypanosoma cruzi via membrane pore formation leading to apoptosis." American Journal of Tropical Medicine and Hygiene **5**(77): 648

R. Mahadeva, R. C. Westerbeek, D. J. Perry, J. U. Lovegrove, D. B. Whitehouse, et al. (1998). "alpha-Antitrypsin deficiency alleles and the Taq-I G -> A allele in cystic fibrosis lung disease." European Respiratory Journal **4**(11): 873-879

W. M. Mars, P. Patmasiriwat, T. Maity, V. Huff, M. M. Weil, et al. (1995). "Inheritance of unequal numbers of the genes encoding the human neutrophil defensins HP-1 and HP-3." J Biol Chem **270**: 30371 - 30376

- H. Matsui, B. R. Grubb, R. Tarran, S. H. Randell, J. T. Gatzky, et al. (1998). "Evidence for periciliary liquid layer depletion, not abnormal ion composition, in the pathogenesis of cystic fibrosis airways disease." Cell **7**(95): 1005-1015
- S. A. McCarroll and D. M. Altshuler (2007). "Copy-number variation and association studies of human disease." Nat Genet **7 Suppl**(39): S37 - 42
- S. A. McCarroll, A. Huett, P. Kuballa, S. D. Cholewicki, A. Landry, et al. (2008). "Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease." Nature Genetics **9**(40): 1107-1112
- H. F. McFarland and R. Martin (2007). "Multiple sclerosis: a complicated picture of autoimmunity." Nature Immunology **9**(8): 913-919
- M. L. Metzker (2010). "Sequencing technologies [mdash] the next generation." Nat Rev Genet **1**(11): 31-46
- K. Miles, D. J. Clarke, W. Y. Lu, Z. Sibinska, P. E. Beaumont, et al. (2009). "Dying and Necrotic Neutrophils Are Anti-Inflammatory Secondary to the Release of alpha-Defensins." Journal of Immunology **3**(183): 2122-2132
- G. Morrison, F. Kilanowski, D. Davidson and J. Dorin (2002). "Characterization of the Mouse Beta Defensin 1, Defb1, Mutant Mouse Model." Infect. Immun. **6**(70): 3053-3060
- M. S. Muhlebach, W. Reed and T. L. Noah (2004). "Quantitative cytokine gene expression in CF airway." Pediatric Pulmonology **5**(37): 393-399
- J. Nathans, T. P. Piantanida, R. L. Eddy, T. B. Shows and D. S. Hogness (1986). "Molecular-Genetics of Inherited Variation in Human Colour-Vision." Science **4747**(232): 203-210
- D. Q. Nguyen, C. Webber, J. Hehir-Kwa, R. Pfundt, J. Veltman, et al. (2008). "Reduced purifying selection prevails over positive selection in human copy number variant evolution." Genome Research **11**(18): 1711-1723
- T. X. Nguyen, A. M. Cole and R. I. Lehrer (2003). "Evolution of primate theta-defensins: a serpentine path to a sweet tooth." Peptides **11**(24): 1647-1654
- F. Niyonsaba, H. Ushio, N. Nakano, W. Ng, K. Sayama, et al. (2007). "Antimicrobial peptides human beta-defensins stimulate epidermal keratinocyte migration, proliferation and production of proinflammatory cytokines and chemokines." Journal of Investigative Dermatology **3**(127): 594-604

- M. Parkes, J. C. Barrett, N. J. Prescott, M. Tremelling, C. A. Anderson, et al. (2007). "Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility." Nature Genetics **7**(39): 830-832
- M. Pazgier, D. M. Hoover, D. Yang, W. Lu and J. Lubkowski (2006). "Human beta-defensins." Cellular and Molecular Life Sciences **11**(63): 1294-1313
- G. H. Perry, N. J. Dominy, K. G. Claw, A. S. Lee, H. Fiegler, et al. (2007). "Diet and the evolution of human amylase gene copy number variation." Nat Genet **10**(39): 1256 - 1260
- G. H. Perry, A. Ben-Dor, A. Tsalenko, N. Sampas, L. Rodriguez-Revenga, et al. (2008). "The Fine-Scale and Complex Architecture of Human Copy-Number Variation." The American Journal of Human Genetics **3**(82): 685-695
- G. H. Perry, F. Yang, T. Marques-Bonet, C. Murphy, T. Fitzgerald, et al. (2008). "Copy number variation and evolution in humans and chimpanzees." Genome Research **11**(18): 1698-1710
- J. Pillay, I. den Braber, N. Vrisekoop, L. M. Kwast, R. J. de Boer, et al. (2010). "In vivo labeling with <sup>(2)</sup>H<sub>2</sub>O reveals a human neutrophil lifespan of 5.4 days." Blood **4**(116): 625-627
- A. Piotrowski, C. E. G. Bruder, R. Andersson, T. D. de Stahl, U. Menzel, et al. (2008). "Somatic mosaicism for copy number variation in differentiated human tissues." Human Mutation **9**(29): 1118-1124
- M. C. Plotkowski, G. Beck, J. M. Tournier, M. Bernardo, E. A. Marques, et al. (1989). "Adherence of *Pseudomonas Aeruginosa* to Respiratory Epithelium and the Effect of Leucocyte Elastase." Journal of Medical Microbiology **4**(30): 285-293
- M. Ramasundara, S. T. Leach, D. A. Lemberg and A. S. Day (2009). "Defensins and inflammation: The role of defensins in inflammatory bowel disease." Journal of Gastroenterology and Hepatology **2**(24): 202-208
- S. Ranganathan, B. Linnane, G. Nolan, C. Gangell and G. Hall (2008). "Early detection of lung disease in children with cystic fibrosis using lung function." Paediatric Respiratory Reviews **3**(9): 160-167
- R. Redon, S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, et al. (2006). "Global variation in copy number in the human genome." Nature **7118**(444): 444 - 454
- J. R. Riordan, J. M. Rommens, B. S. Kerem, N. Alon, R. Rozmahel, et al. (1989). "Identification of the Cystic-Fibrosis Gene - Cloning and Characterization of Complementary-DNA." Science **4922**(245): 1066-1072



R. Sachidanandam, D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein, et al. (2001). "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms." Nature **6822**(409): 928-933

N. H. Salzman, D. Ghosh, K. M. Huttner, Y. Paterson and C. L. Bevins (2003). "Protection against enteric salmonellosis in transgenic mice expressing a human intestinal defensin." Nature **6931**(422): 522-526

B. O. Schroeder, Z. H. Wu, S. Nuding, S. Groscurth, M. Marcinowski, et al. (2011). "Reduction of disulphide bonds unmasks potent antimicrobial activity of human beta-defensin 1." Nature **7330**(469): 419-+

J. Sebat, B. Lakshmi, J. Troge, J. Alexander, J. Young, et al. (2004). "Large-scale copy number polymorphism in the human genome." Science **5683**(305): 525-528

J. Sebat, B. Lakshmi, D. Malhotra, J. Troge, C. Lese-Martin, et al. (2007). "Strong Association of De Novo Copy Number Mutations with Autism." Science **5823**(316): 445-449

M. Sharma, F. Pampinella, C. Nemes, M. Benharouga, J. So, et al. (2004). "Misfolding diverts CFTR from recycling to degradation: quality control at early endosomes." Journal of Cell Biology **6**(164): 923-933

S. Shrestha, J. M. Tang and R. A. Kaslow (2009). "Gene copy number: learning to count past two." Nature Medicine **10**(15): 1127-1129

M. S. Silverberg, J. H. Cho, J. D. Rioux, D. P. B. McGovern, J. Wu, et al. (2009). "Ulcerative colitis-risk loci on chromosomes 1p36 and 12q15 found by genome-wide association study." Nature Genetics **2**(41): 216-220

O. Soehnlein, Y. Kai-Larsen, R. Frithiof, O. E. Sorensen, E. Kenne, et al. (2008). "Neutrophil primary granule proteins HBP and HNP1-3 boost bacterial phagocytosis by human and murine macrophages." Journal of Clinical Investigation **10**(118): 3491-3502

C. C. A. Spencer, Z. Su, P. Donnelly and J. Marchini (2009). "Designing Genome-Wide Association Studies: Sample Size, Power, Imputation, and the Choice of Genotyping Chip." Plos Genetics **5**(5):

M. B. Sporn, A. B. Roberts, L. M. Wakefield and R. K. Assoian (1986). "Transforming Growth-Factor-Beta - Biological Function and Chemical Structure." Science **4763**(233): 532-534

F. Stanke, M. Ballmann, I. Bronsveld, T. Dork, S. Gallati, et al. (2008). "Diversity of the basic defect of homozygous CFTR mutation genotypes in humans." Journal of Medical Genetics **1**(45): 47-54

H. Stefansson, D. Rujescu, S. Cichon, O. P. H. Pietilainen, A. Ingason, et al. (2008). "Large recurrent microdeletions associated with schizophrenia." Nature **7210**(455): 232-U261

Z. M. Sthoeger, S. Bezalel, N. Chapnik, I. Asher and O. Froy (2009). "High alpha-defensin levels in patients with systemic lupus erythematosus." Immunology **1**(127): 116-122

B. E. Stranger, M. S. Forrest, M. Dunning, C. E. Ingle, C. Beazley, et al. (2007). "Relative impact of nucleotide and copy number variation on gene expression phenotypes." Science **5813**(315): 848-853

P. H. Sudmant, J. O. Kitzman, F. Antonacci, C. Alkan, M. Malig, et al. (2010). "Diversity of Human Copy Number Variation and Multicopy Genes." Science **6004**(330): 641-646

M. A. Taub, H. Corrada Bravo and R. A. Irizarry (2010). "Overcoming bias and systematic errors in next generation sequencing data." Genome medicine **12**(2): 87

R. Tirouvanziam, I. Khazaal and B. Peault (2002). "Primary inflammation in human cystic fibrosis small airways." American Journal of Physiology-Lung Cellular and Molecular Physiology **2**(283): L445-L451

Y. Tsutsumi-Ishii, T. Hasebe and I. Nagaoka (2000). "Role of CCAAT/enhancer-binding protein site in transcription of human neutrophil peptide-1 and -3 defensin genes." J Immunol **6**(164): 3264-3273.

E. V. Valore and T. Ganz (1992). "Posttranslational processing of defensins in immature human myeloid cells." Blood **6**(79): 1538-1544

C. K. van der Ent and P. L. P. Brand (1999). "Advantages of comprehensive lung function techniques in cystic fibrosis." Journal of the Royal Society of Medicine **92**: 13-18

L. L. Vanscoy, S. M. Blackman, J. M. Collaco, A. Bowers, T. Lai, et al. (2007). "Heritability of lung disease severity in cystic fibrosis." American Journal of Respiratory and Critical Care Medicine **10**(175): 1036-1043

J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, et al. (2001). "The sequence of the human genome." Science **291**: 1304 - 1351

- D. Vollrath, J. Nathans and R. W. Davis (1988). "Tandem Array of Human Visual Pigment Genes at Xq28." Science **4859**(240): 1669-1672
- S. Vylkova, N. Nayyar, W. S. Li and M. Edgerton (2007). "Human beta-defensins kill *Candida albicans* in an energy-dependent and salt-sensitive manner without causing membrane disruption." Antimicrobial Agents and Chemotherapy **1**(51): 154-161
- S. Walker, S. Janyakhantikul and J. A. L. Armour (2009). "Multiplex Paralogue Ratio Tests for accurate measurement of multiallelic CNVs." Genomics **1**(93): 98-103
- T. Walsh, J. M. McClellan, S. E. McCarthy, A. M. Addington, S. B. Pierce, et al. (2008). "Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia." Science **5875**(320): 539-543
- C. Webber, J. Y. Hehir-Kwa, D. Q. Nguyen, B. B. A. de Vries, J. A. Veltman, et al. (2009). "Forging Links between Human Mental Retardation-Associated CNVs and Mouse Gene Knockout Models." Plos Genetics **6**(5):
- W. C. Wimley, M. E. Selsted and S. H. White (1994). "Interactions between Human Defensins and Lipid Bilayers- Evidence for Formation of Multimeric Pores." Protein Science **9**(3): 1362-1373
- F. A. Wright, L. J. Strug, V. K. Doshi, C. W. Commander, S. M. Blackman, et al. (2011). "Genome-wide association and linkage identify modifier loci of lung disease severity in cystic fibrosis at 11p13 and 20q13.2." Nature Genetics **6**(43): 539-U567
- A. Xaubet, A. Marin-Arguedas, S. Lario, J. Ancochea, F. Morell, et al. (2003). "Transforming growth factor-beta(1) gene polymorphisms are associated with disease progression in idiopathic pulmonary fibrosis." American Journal of Respiratory and Critical Care Medicine **4**(168): 431-435
- Y. J. Xiao, A. L. Hughes, J. Ando, Y. Matsuda, J. F. Cheng, et al. (2004). "A genome-wide screen identifies a single beta-defensin gene cluster in the chicken: implications for the origin and evolution of mammalian defensins." Bmc Genomics **5**):
- Y. L. Xue, Q. J. Wang, Q. Long, B. L. Ng, H. Swerdlow, et al. (2009). "Human Y Chromosome Base-Substitution Mutation Rate Measured by Direct Sequencing in a Deep-Rooting Pedigree." Current Biology **17**(19): 1453-1457
- E. Zeggini, L. J. Scott, R. Saxena, B. F. Voight, J. L. Marchini, et al. (2008). "Meta-analysis of genome-wide association data and large-scale replication

identifies additional susceptibility loci for type 2 diabetes." Nature Genetics **5**(40): 638-645

G. Z. Zou, E. de Leeuw, C. Li, M. Pazgier, C. Q. Li, et al. (2007). "Toward understanding the cationicity of defensins - Arg and Lys versus their noncoded analogs." Journal of Biological Chemistry **27**(282): 19653-19665